

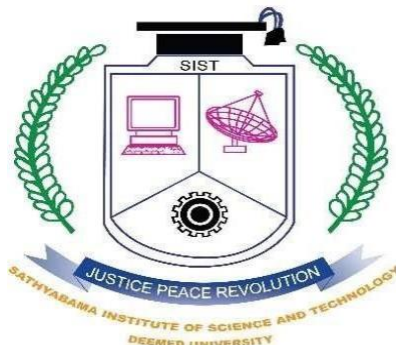
**TRANSCRIPTOME WIDE ASSOCIATION STUDIES
FOR ALZHEIMER'S DISEASE**

Submitted in partial fulfillment of the requirements for the
award of Bachelor of Technology Degree in
BIOINFORMATICS

By

SRAVANI VALLIGARI (Register No.37220022)

HEMA KIRAN .M.B (Register No.37220007)



**DEPARTMENT OF BIOINFORMATICS
SCHOOL OF BIO AND CHEMICAL ENGINEERING
SATHYABAMA**

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

**Accredited with Grade "A" by NAAC
JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI – 600119
MARCH 2021**



SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade "A" by NAAC
JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119
www.sathyabama.ac.in



BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **SRAVANI VALLIGARI (Register No.37220022)** and **HEMA KIRAN .M.B (Register No.37220007)** who carried out the project entitled **"TRANSCRIPTOME WIDE ASSOCIATION STUDIES FOR ALZHEIMER'S DISEASE"** under my supervision from December 2020 to March 2021.

Internal Guide

Dr. DANIEL ALEX ANAND

Head of the Department

Dr. JEMMY CHRISTY

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

We **SRAVANI VALLIGARI (Register No.37220022)** and **HEMA KIRAN .M.B (Register No.37220007)** hereby declare that the Project Report entitled **“TRANSCRIPTOME WIDE ASSOCIATION STUDIESFOR ALZHEIMER’S DISEASE”** done by us under the guidance of **Dr. DANIEL ALEX ANAND** is submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering degree in Bioinformatics.

DATE:

PLACE: CHENNAI

SIGNATURE OF THE CANDIDATES

ACKNOWLEDGEMENT

We are pleased to acknowledge our sincere thanks to Board of Management of **SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY** for their kind encouragement in doing this project and for completing it successfully. We are grateful to them.

We convey our thanks to **Dr. JEMMY CHRISTY, Head of the Department, Dept. of Bioinformatics** for providing us necessary support and details at the right time during the progressive reviews.

We would like to express our sincere and deep sense of gratitude to our Project Guide **Dr. DANIEL ALEX ANAND, Dept of Bioinformatics** for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of our project work.

We wish to express our thanks to all Teaching and Non-teaching staff members of the **Department of Bioinformatics** who were helpful in many ways for the completion of the project.

ABSTRACT

The most common form of dementia, Alzheimer's disease, is becoming prevalent day after day. There are many causes behind the progression of the disease; it could be genetic, hereditary or environmental conditions. The exact mechanism that is involved in the progression of the disease remains unresolved. When it comes to the genetics of the disease, it is assumed that the genes that are expressed during disease progression are responsible for the disease's occurrence. However, it is important to note that certain genes are silent mutations, meaning they are present in the disease-causing genome or transcriptome but are not expressing in any way that affects the disease. These genes are considered as false positives, and they are also found by certain novel genes. To find these genes, we took genome and transcriptome samples from 64-years old women with Braak stage III and Braak stage VI AD in the frontal cortex, as well as post-mortem brain tissue from controls and AD patients. In this study, we have summarized on the GWAS and transcriptome studies to identify and analyse single nucleotide polymorphisms (SNPs) in the genome and Differentially Expressed Genes (DEGs) in the transcriptome of Alzheimer's disease. The up-regulated and down-regulated genes are analyzed between the AD Vs Non-AD individuals. We have identified a total of 46 genes that are expressed in the transcriptome in which all are novel genes and some are false positives. Genome Asia 100k was performed and analysed to determine the population distribution of allele frequencies of SNIPS in the specified population. By using PhenGenI and IGAP 82 SNPs were identified which are present in and expressed in the phenotype of persons with AD. Using GWAS studies total of 60 SNPs were predicted which are present in the genome of AD. Further research is underway in the construction of a network for identified DEGs and the development of a pipeline through the integration of AD's genome and transcriptome. This study can help to understand the mechanism of protein and can help to detect drugs.

Key words: Alzheimer's disease, transcriptome, genome, novel genes, false positives, phenotype, allele frequency, SNPs, DEGs.

INDEX

S.NO.	TITLE	PAGE NO.
1	LIST OF FIGURES	IX
2	LIST OF TABLES	XI
3	LIST OF ABBREVIATIONS	XII
4	CHAPTER 1 – INTRODUCTION 1.1. OUTLINE OF THE PROJECT 1.1.1. Alzheimer’s disease 1.1.2. Genes those are involving in the disease progression of AD 1.1.3. Integrating genome and transcriptome of AD 1.2. POPULATION GENETICS 1.3. OUT SOURCE 1.3.1. Tools and Software that are used	1-7
5	CHAPTER 2 – LITERATURE SUVEY 2.1 REVIEW – LITERATURE WORK 2.1.1 What is Alzheimer’s disease and the pathophysiology of this disease? 2.1.2 How do the genes in AD play a significant role? 2.1.3 What is the role of AD’s genome and transcriptome?	8-18
6	CHAPTER 3 – AIM AND OBJECTIVE	19
7	CHAPTER 4 – MATERIALS AND METHODS 4.1 MATERIALS THAT ARE USED 4.1.1 Sample Collection 4.1.1.1 NCBI 4.1.1.2 SRA 4.1.2 Faster download and extract reads in FASTA format from NCBI SRA 4.1.3 FASTQC Read Quality Report 4.1.4 Trimmomatic 4.1.5 HISAT2 4.1.6 CuffLink 4.1.7 CuffMerge	20-35

	4.1.8 Cuffdiff 4.1.9 Bowtie2 4.1.10 SAM 4.1.11 Variant Calling 4.1.12 Expression Atlas 4.1.13 PhenGenI 4.1.14 GO Annotation 4.1.15 Genome Asia 100k 4.1.16 IGAP 4.1.17 Manhattan Plot 4.1.18 Volcano Plot 4.1.19 DEG HeatMap 4.2 METHDOLOGY 4.2.1 Sample 4.2.2 Extract reads from FASTQ (Primary Analysis) 4.2.3 Quality check 4.2.4 Trimming 4.2.5 Alignment (Secondary Analysis) 4.2.5.1 Bowtie2 4.2.5.2 HISAT2 4.2.6 SAMtools 4.2.7 Variant Calling 4.2.8 SNP Annotation 4.2.9 Transcript Assembly 4.2.9.1 Cufflinks 4.2.9.2 Cuffmerge 4.2.9.3 Cuffdiff 4.2.10 Identifying SNPs in genotype and phenotype 4.2.11 Population Genetics 4.2.12 GWAS Data	
8	CHAPTER 5 – RESULTS AND DISCUSSION 5.1 QUALITY CHECK REPORT OF THE RAW SEQUENCES 5.1.1 Base statistics of raw sequence 5.1.2 Sequence quality check 5.1.3 Sequence length distribution 5.2 ALIGNMENT RESULTS 5.2.1 Bowtie2 alignment results 5.2.2 HISAT2 5.3 SAMTOOLS RESULTS 5.4 VARIANT CALLING 5.5 ASSEMBLED READS 5.5.1 Cufflinks 5.5.2 Cuffmerge	35-65

	5.6 DIFFERENTIALLY REGULATING GENES AND BASE STATISTICS 5.7 SNPs THAT ARE EXPRESSED IN BOTH GENOTYPE AND PHENOTYPE 5.8 GO ANNOTATION 5.9 VOLCANO PLOT 5.10 HEAT MAP 5.11 BOX PLOT 5.12 DENSITY PLOT 5.13 HISTOGRAM 5.14 MEAN VARIANCE 5.15 POPULATION GENETICS 5.16 GWAS DATA 5.17 MANHATTAN PLOT	
9	CHAPTER 6 – SUMMARY AND CONCLUSION 6.1 SUMMARY 6.2 CONCLUSION AND FUTURE WORK	66
10	REFERENCES	67-69

LIST OF FIGURES

FIG NO.	TITLE	PAGE NO.
1.1	Various Hypothesis of Alzheimer's disease in clinical trials upto 2019	2
4.1	Sample of genome dataset collection in SRA Run Selector	21
4.2	Sample of transcriptome data collection in SRA Run Selector	22
4.3	Expression Atlas for DEG	25
4.4	Overall Workflow in finding the differentially Expressed Genes	28
4.5	PhenGenI Summary results for Alzheimer's Disease	33
4.6	PhenGenI Association Table for Alzheimer's Disease	34
4.7	IGAP meta-analysis data of GWAS data in Alzheimer's disease	34
5.1	Base Statistics of the raw sequence of run ERR1378585	36
5.2	Base Statistics of the raw sequence of run SRR12828917	37
5.3	Per base sequence quality report of raw sequence of run ERR1378585	38
5.4	Per base sequence quality report of raw sequence of run SRR12828917	38
5.5	Sequence length distribution of the run ERR1378585	39
5.6	Sequence length distribution of the run SRR12828917	39
5.7	Alignment of raw sequence (genome) with the reference genome hg38	40
5.8	Bowtie2 alignment statistics of ERR1378585	40
5.9	HISAT2 alignment report of raw sequence (transcriptome) with reference genome hg38	41
5.10	HISAT2 alignment statistics of ERR1378585	41
5.11	SAMtools filtered alignments of genome (BAM)	42
5.12	Call Variants of genome (VCF)	42
5.13	Snpsift annotation of genome (rs ID's of SNPs)	43
5.14	Cufflinks – the transcripts are assembled	43
5.15	Merged transcriptome	44
5.16	GO component	54
5.17	GO Process	54
5.18	GO function	55
5.19	Up-regulated & Down-regulated genes in AD Vs Non-Ad	56
5.20	Clustering of expressed genes based on gender	56
5.21	Clustering of expressed genes based on disease state	57
5.22	Clustering of expressed genes based on tissue	57
5.23	Gene Expression comparison between AD Vs Non-AD	58

5.24	Expression rate between AD Vs Non-AD	59
5.25	P-value distribution	59
5.26	Mean-Variance	60
5.27	SNP distribution of population – 100k genome	60
5.28	SNP distribution in Manhattan Plot	65

LIST OF TABLES

S.NO.	TITLE	PAGE NO.
2.1	Alzheimer's disease most common genes	10
5.1	Differentially regulating genes	45-52
5.2	List of SNPs (rs ID's) expressed in genotype and phenotype of AD	53
5.3	SNPs identification through GWAS	61-64

LIST OF ABBREVIATIONS

AD	Alzheimer's disease
IGAP	International Genomics of Alzheimer's project
GWAS	Genome Wide Association Studies
NGS	Next Generation Sequencing
PhenGenI	Phenotype – Genotype Integrator
DEG	Differently Expressed Genes
SNP	Single Nucleotide Polymorphism
APP	Amyloid Precursor Protein
APOE	Apolipoprotein E
LOAD	Late onset Alzheimer's disease
EOAD	Early onset of Alzheimer's disease
DMTs	Disease modifying therapies of the agent
PSEN1	Presenilin 1
PSEN2	Presenilin 2
CNV	Copy Number Variation
PRS	Polygenic risk score
PHS	Polygenic hazard score
SAD	Sporadic later Alzheimer's disease
FAD	Family Alzheimer's disease
FTD	Front temporal dementia
eQTLs	Expression quantitative trait loci
NCBI	National Centre for Biotechnology Information

SRA	Sequence Read Archive
SAM	Sequence Alignment Map
NHI	National Institutes of Health

CHAPTER 1

INTRODUCTION

1.1. OUTLINE OF THE PROJECT

1.1.1 *Alzheimer's Disease*

Dr. Alois Alzheimer is the person who gave his name to the disorder. Dr. Alzheimer found differences in a woman's brain tissue in 1906 who had died of an unusual mental condition. Her symptoms included memory loss, language difficulties, and behavior that was erratic. He studied her brain after she died and noticed several irregular clumps (now known as amyloid plaques) and twisted bundles of fibres (now called neurofibrillary, or tau, tangles).

Alzheimer's disease is an irreversible, progressive brain disease that gradually erodes memory and thinking skills, as well as the ability to perform even the most basic tasks. Symptoms appear in the mid-60s in the majority of people with the disease that is with late-onset AD. Early-onset Alzheimer's disease is extremely rare and occurs between the ages of 30 and 60. The most common cause of dementia among older adults is Alzheimer's disease. The resulting signs of billions of nerve cells or neurons in the brain are the memories, thinking, communication or feelings of a healthy human.[1] These neurons communicate constantly through the active passing of electrical charges which cause chemical (neurotransmitters) to be released across other neurons.

Toxic changes in the brain destroy this healthy balance in people with Alzheimer's disease. Two major proteins beta-amyloid and Tau are involved in this disease.

Amyloid consists of the breakdown of a larger amyloid protein precursor. This protein's abnormal levels clump in and form plaques between neurons that disrupt its function.[1,2] Tau usually attaches to and stabilizes microtubules in healthy neurons. However, abnormal changes in chemistry in Alzheimer's disease cause tau to detach itself from microtubules and stick to additional tau molecules to form threads that finally become enmeshes within the neurons.

These tangles block the transportation system of the neuron which harms synaptic neuron communication. Not only are tau and beta-amyloid factors involved in Alzheimer's disease, there are many other factors that are involved in the cause of the disease. The vascular system may fail to supply the brain with sufficient quantities of blood and nutrients. It's possible that the brain doesn't have enough glucose to keep going.

Due to chronic inflammation inside the brain, Microglia, the cell that removes the debris, may fail to do its job. Neurons lose their ability to communicate over time, and as they die, the brain shrinks.[3]

As more neurons die, the individual will gradually lose the ability to independently think, recognize, make choices, communicate and function.

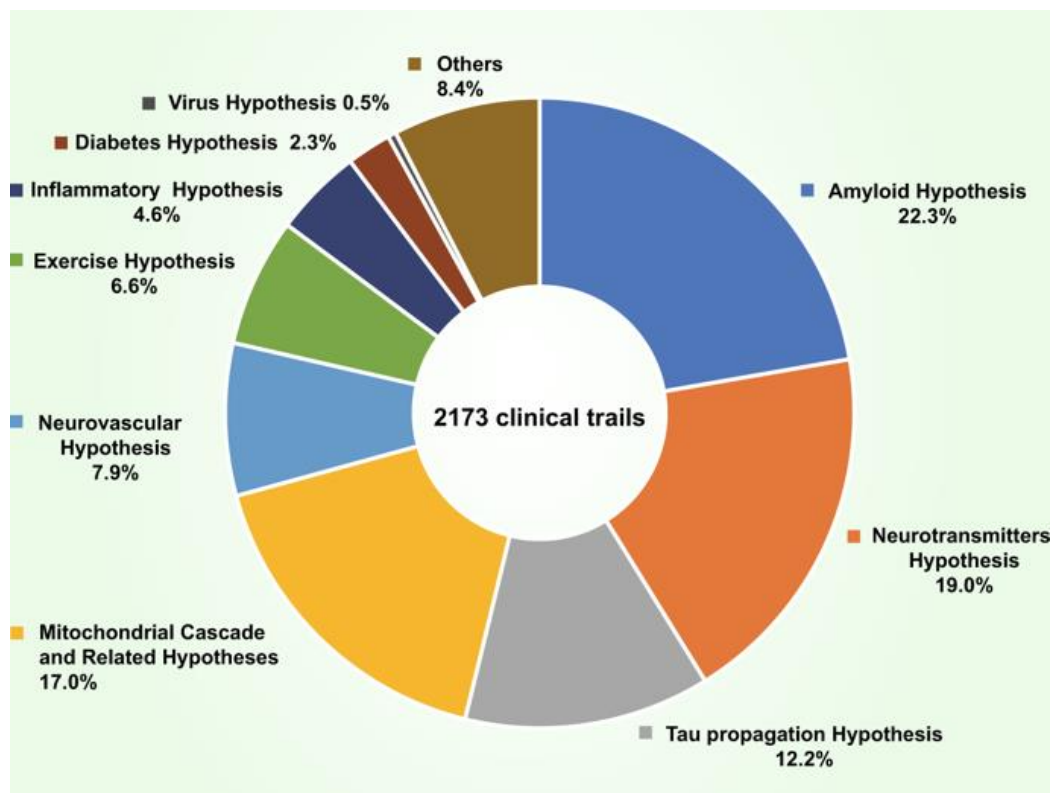


Fig:1.1 Various Hypotheses of Alzheimer's disease in Clinical Trials up to 2019
(Liu, Pei-Pei, et al. 2019)

1.1.2 Genes those are involving in the disease progression of AD.

- Three genes (APP, PSEN1, and PSEN2, which code for amyloid precursor protein, presenilin-1, and presenilin-2) have been identified as being responsible for early onset, dominantly inherited forms of AD.[4]
- These rare mutations, which caused the Autosomal-dominant forms of the disease, gave rise to an Amyloid cascade hypothesis that radically changed our understanding of AD.
- The e4 allele of the apolipoprotein E (APOE) gene was linked to the risk of developing Alzheimer's disease in 1993.
- More than 40 genes/loci have been linked to the risk of Alzheimer's disease in the last ten years.
- Despite the fact that some of these signals are likely to be false positives, genetic knowledge has shed new light on the pathogenesis of Alzheimer's disease.[4]

1.1.3 Integrating Genome and Transcriptome of AD

MULTI OMICS - an approach in which data sets of different omic groups are combined during analysis. To build network between them, we use a combination of Genome and Transcriptome datasets.

Genomic - Genome wide Association study (GWAS): An approach to finding genetic changes related to the specific disease, by rapidly scanning markings across complete sets of DNA or genomes. New genetic associations can be used to develop better strategies for the detection, treatment and prevention of disease.

Transcriptome - Transcriptome is a well-known advanced gene expression protocol and a novel RNA discovery. The analysis of gene expression can provide a snapshot, under different conditions, of active genes and transcripts

Highly advanced genomic approaches shape our understanding of Alzheimer's disease genetics. The field AD GWAS moved in a couple of additional directions and resulted in the discovery of additional risk factors in AD genetics.[4] The discovery stage (e.g. N=17 008 cases of AD and 37 154 controls analyzed in the first IGAP study) of the GWAS

consortia (IGAP) has increased considerably and its statistical strength has been improved.

Next Generation Sequencing techniques enabled the detection of large number of rare variants that requires larger number of samples. These sequencing approaches has shown the rare variants in TREM2, SORL1 and ABCA7 for example are major genetic risk factors for AD.[4]

Whole-Exome Sequencing Technique took place in the largest sequencing effort in AD to date showing results for 5740 cases of late-onset AD and 5096 cognitively normal controls. Functional variants and the pathways in which the genes are potentially involved can be identified by combining quantitative expression of loci, active enhancers and/or genes indifferent cell types and disease tissues expression. Post genomic studies play the most important role in providing the meaningful information about the characterization of missing genetic risk factors for AD.

Pathophysiology of AD is a challenging area that requires deep understanding about the mechanism involved in the post genomic area of AD. Change in the gene expression and gene regulation might show strong impact in the in the Neuro degeneration.[5]

A Transcriptome study reveals the importance of analysis of gene expression pattern and provides insight into the mechanism underlying complex traits in humans. A number of variants and alternative splicing forms may reduce the risk of neurodegeneration. For example, a variant on the HMG-CoA reductase (HMGCR) enzyme gene rs3846662 could cause exon 13 to be spliced differently and deleted.[5]There are various methods that could be used for the Transcriptome analysis.

In any method the first step need to do is RNA extraction, removal of RNA from DNA, RNA protein complexes, DNA removal by DNase. The genetic–environmental relationship may be discovered using gene expression profiling of blood and brain tissue samples. It's been suggested that differentially expressed genes and non-coding RNAs play a role in disease pathogenesis.

Tissue samples can be taken from blood, plasma, serum, brain cells, cerebrospinal fluid and other parts of the brain. The brain tissues are the most effective markers for the disease-related changes in Transcriptome. The techniques that can be used in Transcriptome studies are Array Techniques to identify the DEGs, splicing, miRNA analysis, CNV. RNAseq technique such as Next Generation Sequencing(NGS) in detection of DEGs, splicing, miRNA analysis, CNV. qRT-PCR technique to find the DEGs, miRNA. Isothermal amplification technique for miRNA detection.

The DEG analysis was performed to find the up regulated and down regulated genes are present in the disease compared with unaffected controls. The over expressed or under expressed of different genes may result in different consequences like immune system imbalance, metabolite alterations and disease formation. There are many sequencing tools to perform RNAseq such as Roche 454, Illumina, Helicos (DNA polymerase based), or PacBio, SoLid (ligase based). It is important to see if transcriptions can lead to additive effects in the known risk factors for diseases and disease-related pathways.[5]

Underlying the role of Genome and Transcriptome of AD identified the novel genes from the Differently Regulating Genes with the false positives. These novel genes are expressed in great amount in the post-mortem brain tissue of controls and Alzheimer's disease patients with different APOE genotypes.

The GWAS studies on the genome of Single-cell whole genome sequencing of brain cells of Alzheimer's disease patients identified the most common SNIPS that are allocated between the IGAP and the SNIPS of the sample genome with 180 common SNIPS.

The up-regulated and down-regulated expression data from the post-mortem brain tissue of controls and Alzheimer's disease patients with different APOE genotypes shows that the DEGs are most significant with $P_{adj} < 0.5$.

1.2 POPULATION GENETICS

Darwin's Origin of Species was published in 1859 to present two major theses. 1) Modern humans were inherited from common ancestors 2) the natural selection process was the main evolutionary change mechanism.

The first thesis was easily accepted by all, but they refused to agree that evolution took place through natural selection. Later on, the inheritance theory of Mendel made a straightforward declaration that an organism's phenotype requires inheritance. Mendel believed that the majority of phenotypic characteristics were controlled by a single pair of factors, such as the shape of seeds in his pea plants, but now it is understood that several pairs of Genes, not just one, influence most characteristics.[6]

In our interpretation of inheritance, this hypothesis marks a turning point. Single Nucleotide Polymorphisms (SNPs) are the most common form of genetic variation that leads to changes in an organism's phenotype. These SNPs are involved directly or indirectly in the movement of genes.

Gene Flow is the mechanism of allele shift passing from one population to another. The Phenotypic difference can be height, form, colour, weight between species. All of these features are regulated by genes. Trying to explain and measure human phenotypic variation is complicated because of the complex relationship between genes and the environment, as well as between multiple genes. The nucleotide sequence of the genome, which may speak of the particular human variation, is therefore very easy to take. This can be a way of defining the SNPs in the identified population of the person.

Genetic drift is a phase of evolution in which a population's allele frequencies change over generations due to chance (sampling error). To be more accurate, due to "sampling error" in choosing the alleles from the gene pool of the present generation for the next generation, genetic drift is altered. While genetic drift occurs in populations of all sizes, in small populations, its effects appear to be greater. Due to stochastic sampling error, genetic drift occurs more often in small samples.

The SNP distribution of AD in the Indian population can be analyzed by taking the SNIPS of the genome from the Single-cell whole genome sequencing of brain cells of Alzheimer's disease patients using the Genome Asia 100k.

The SNPs spread of the population frequency can be identified for each variant. It shows the allele count across the population with different statistics varying from variant to variant.

1.3. OUT SOURCE

1.3.1 Tools and Software's that are used

The software that was used is Python 3 for the DEGs prediction, to identify the up regulated and down regulated genes in the Transcriptome of AD.

The IGAP (International Genomics of Alzheimer's project) that contains the summary statistics of the meta-analysis of the genome-wide studies of AD was used to find the SNIPS that are in common between sample genome and the IGAP.[4]

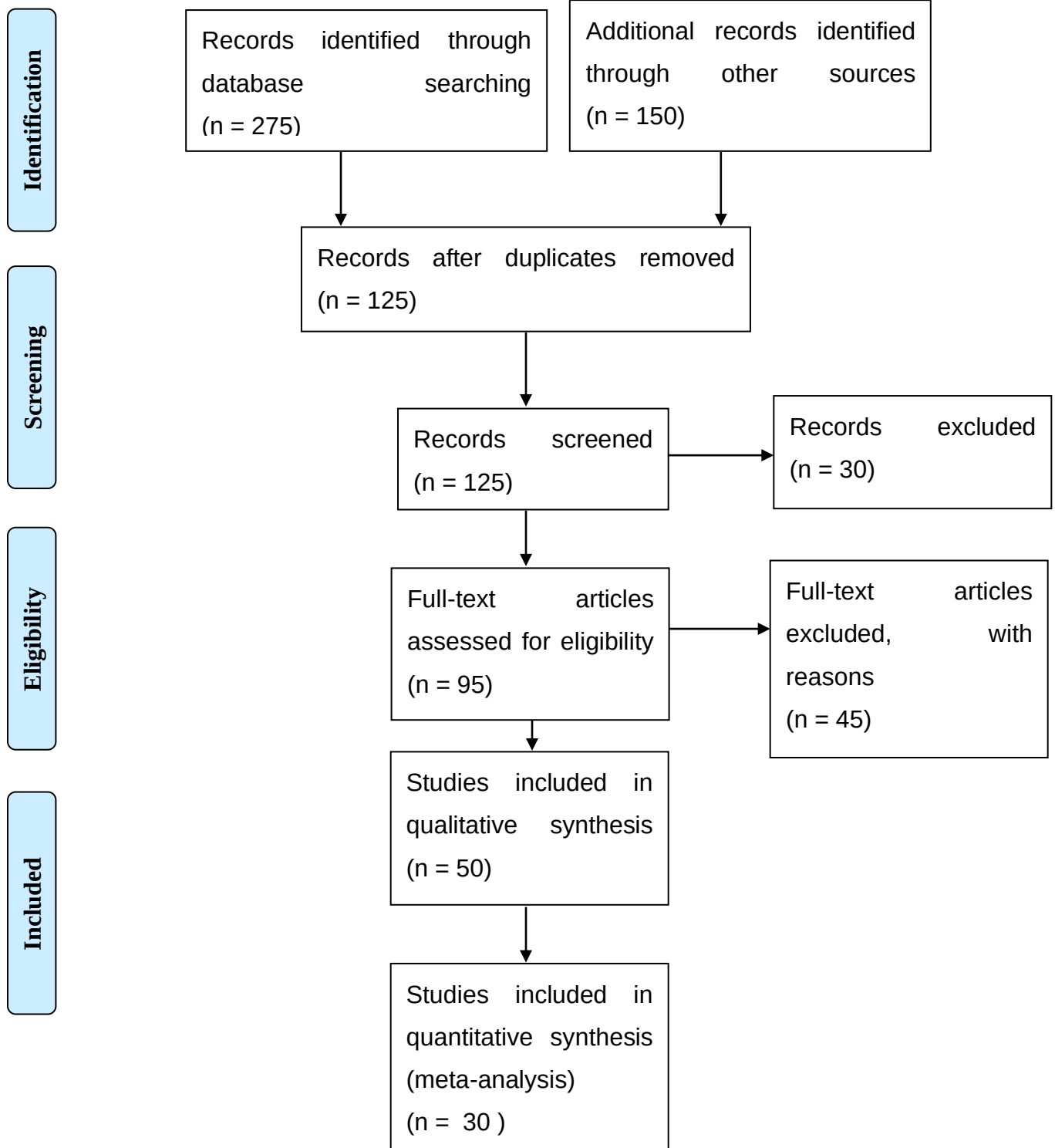
IGAP has 2 datasets, one is the meta-analysis results of stage 1 of genotype and imputed data and second one contains the meta-analysis results of genotype and independent sets of AD patients with combined stage1/stage 2 P values.

PhenGenI - Phenotype and Genotype Integrator was used which contains the data of the phenotypic trait of AD with their association results, SNPs, Genes etc.

Genome Asia 100k project that maintains the information related to the population wide genome, allele frequencies based on different search fields like rs ID, variant, gene, location.

CHAPTER 2

LITERATURE SURVEY



2.1. REVIEW-LITERATURE WORK:

2.1.1 What is Alzheimer's disease and the pathophysiology of this disease?

Alzheimer's disease is a chronic illness that degenerates and destroys brain cells steadily. This induces the early onset of thought loss called Dementia AD may evolve for a variety of reasons, such as environmental lifestyles, Genetic AD arises mainly because of genetic modifications, and both inherited and genetic disorders may be the cause of Alzheimer's disease

Genetic mutations that are passed from parent to child are the most common cause. Basic gene mutations and multi gene variants may also be. In both of these mutations, the genes involved are different.

It is shown, according to previous reports that mutations have mainly occurred within a single gene that causes the disease. In such cases, for several generations, the mutations are passed from parent to child.

People with mutations of this kind appear to get the disease in their 30s, 40s, and 50s, which is expected to happen in the late 60s or 65s. It is considered Early-onset dementia for those who develop the disorder in their 60s or 65s. Many who experience the condition after their 65s or 70s are similarly referred to as late-onset dementia.

AD pathophysiology is difficult as there is no proper medication to cure the disease and the cause of the disease is not completely known and the essence of the disease is not properly understood. Intracellular neurofibrillary tangles[1,7], which is one of the characteristics of AD that includes twisting the 'Tau' protein and senile plaques that arise when amyloid beta ($A\beta$) protein misfolds, can be depicted in disease pathophysiology.

The development of AD can currently be decreased by taking some of the drugs prescribed, such as donepezil, rivastigmine, galantamine and meantime. On-going research focuses primarily on how to avoid the progression of the disease by establishing disease modifying therapies of the agent (DMTs)[7]

2.1.2 How do the genes in AD play a significant role?

A mutation in one of only three genes normally triggers Alzheimer's disease.

1. on chromosome 14, Presenilin 1 (PSEN1)
2. on chromosome 11, Presenilin 2 (PSEN2)
3. on chromosome 21, the amyloid precursor protein (APP)

The early-onset of Alzheimer's disease (EOAD) (30s-50s) includes these genes.

The development of abnormal proteins occurs when one of these three genes undergoes mutation. These mutations will include those resulting in the development of amyloid plaques in the breakdown of APP.

The genes associated with Alzheimer's late onset are Apolipoprotein E. (APOE).[8]

There are 3 types of alleles in Apolipoprotein E (APOE)

1. ApoE e2 tends to decrease the risk of Alzheimer's, the least common
2. ApoE e3
3. The risk of Alzheimer's is raised by ApoE e4, which is a little more normal.

Table 2.1 Alzheimer Disease (AD) most common Genes

The genes of Alzheimer's disease are found on 4 distinct chromosomes and are associated with sporadic or autosomal dominant ancestry. [8]

AD LOCI	GENE SYMBOL	GENE NAME	CHROMOSOME	INHERITANCE
AD1	APP	Amyloid precursor protein	21q21	Autosomal Dominant
AD2	APOE	Apolipoprotein E	19q13.32	Sporadic
AD3	PSEN1	Presenilin 1	14q24.2	Autosomal Dominant

AD4	PSEN2	Presenilin 2	1q42.13	Autosomal Dominant
-----	-------	--------------	---------	-----------------------

So far, we have an idea of how the disease happens, what all the genes are involved in the development of the disease. Now, let's see how these genes are involved in the development and mechanism of the disease.

The Amyloid precursor protein (APP) gene is translated by β - and γ -secretase into A β peptides. Increased self-aggregation of A β peptides or increased development of A β is identified as pathogenic mutations in APP.

The genes PSEN1 and PSEN2 encode part of the γ -secretase complex and most of the identified mutations in autosomal dominant AD are accounted for by PSEN1. The PSEN1 pathogenic mutation is involved in reducing both A β 42 and A β 40 development.

AD, the A β plays a significant role. The mutated PSEN1 decreases the development of A β , which includes memory loss and the role of nerve cells. [9]

As a significant genetic risk factor for LOAD (Late onset Alzheimer's disease), the APOE4 gene that has APOE4 allele has been identified. LOAD has been described as the most common AD that contains a high number of uncommon risk variants. [9] In order to get the right results from the false-positives, more experiments need to be performed. The allele frequencies of rare variants differ from population to population, making it difficult to study.

ABCA7, CD33, CLU, CR1, MS4A, INPP5D, TREM2, PLCG2, PTK2B, and ABI3 are the risk genes involved in inflammatory response; APOE, CLU, ABCA7, and PLCG2 are involved in lipid metabolism; BIN1, PICALM, CD2AP, EPHA1, and SORL1 are involved in endocytosis/vesicle-mediated transport; RANBP2 is involved in cell cycle regulation; MEF2C is involved in oxidative stress response; and UNC5C is involved in axon guidance. [9]

2.1.3 What is the role of AD's genome and transcriptome?

More than 20 disease-associated genomic loci have been identified in recent studies, but there are several undiscovered loci. Study of the genome-wide relationship between clinically diagnosed AD and AD-by-proxy (71,880 cases, 383,378 controls). A strong genetic association of $rg = 0.81$ was found between AD and AD-by-proxy.

The risk genes and the possible causative genes involved in immune-related tissues and cell types can be identified through meta-analysis. Information on the biological pathways involved in the degradation of amyloid precursor proteins and lipid-related processes was provided in the Gene-set study. Further research is underway to identify the genetic factors leading to the risk of AD and to provide new insights into AD neurobiology. [10]

With several environmental and genetic variables, AD has a complex etiology. In order to find that common genetic markers play a major role in AD vulnerability and development, the Genome-wide association analysis was performed. In order to find the genetic loci related to late-onset AD susceptibility, the genes are repeated. In order to alter the variations in gene size, pathway analysis was performed and the association structures between the SNPs were retained.

This is accomplished by gene enrichment and by contrasting this with null hypothesis distribution dependent on random permutation. [11] AD susceptibility variants reported may have limited utility in predicting disease progression in patients with AD. It is possible to use disease progression for further GWAS studies, and deep sequencing studies may improve the understanding of the AD mechanism.

GWAS allows variables that quantify genetic responsibility for a disease or trait to understand the genotype data. The analysis of the polygenic risk score (PRS) and the analysis of the polygenic hazard score (PHS) are two methodological methods compared to recent GWAS data and have shown that they have achieved similar results. In order to find the number of risk alleles in an individual, the PRS is used. TREM2 is recognized by the use of exome sequencing, which has become increasingly evident in the importance of microglia. [12] One factor in assessing one's risk of

developing illness was the genetic difference in microglial response to A β . Further understanding in the pathogenic process will improve in the effective treatment and diagnosis of the disease.

Several studies have shown that improvements in gene expression and gene regulation can have a severe effect on the progression of the disease. Various factors can play an important role in disease onset, such as mRNA-transcription factor interactions, non-coding RNAs, alternate splicing, or copy number variants. [13] The for better care and diagnosis of the disease, it is important to recognize the transcriptome of AD.

The genetic risk factors at the transcriptome level can assist in the discovery of multifactorial AD genetic and transcriptome investigations. Additional information on disease-associated mechanisms can also be given to research additional factors, such as gene-expression modifications, non-coding RNAs, CNVs, or splicing. Modifications to gene expression may help prevent disease, some miRNAs may be beneficial for neurons, and some may be involved in neurodegeneration. [13]

For AD therapies, there are some miRNA candidates that were evaluated in cell. MiR-455-3p, which has been detected in neuroblastoma cells, plays an important role in neuroprotection by reducing mutant APP expression and A β 42 and A β 40 peptide levels. Using Roche 454, Illumina, Helicos, PacBio, Strong, nanopore sequencing methods, the RNAseq system is used to find the DEGs, Splicing, miRNA analysis and CNV (Copy Number Variation)

The benefit of the RNAseq experiment is that the proceeds are clear with a simple workflow and the larger scale of variants is also easy to detect.[13] The DEG miRNA can be identified using Applied BioSystems, BioRAD methods, using the qRT-PCR technique.

In AD, there are several downregulated genes involved in the encoding of metastable proteins involved in homeostasis, resulting in irregular amyloid deposits, phosphorylation of Tau, or inflammation. In order to encounter the progression of the disease, pathway development of gene expression is mandatory and also helps in the better diagnosis of AD. To get the precise data on DEG (Differentially Expressed

Genes), alternative splicing, or miRNAs, microarrays and Next Generation Sequencing(NGS) techniques can be used.

In the analysis of transcriptome, there are many limitations and difficulties, such as the consistency of RNA can be difficult to maintain, the transcriptome profiles may vary from tissue to tissue. Without much trouble having AD biomarkers and mild cognitive disability, peripheral samples such as peripheral blood, serum, plasma, or peripheral blood mononuclear cells can easily be taken. Studying the transcriptome nature of the blood in the brain may help to better understand AD, but there may be a risk that blood does not express several proteins. In order to find out how gene expression alteration can be successful in the human brain, more studies are needed.

The use of preclinical studies with human-induced pluripotent stem cell technologies (iPSC's) and use of iPSC-based cell types for the study of a range of human conditions including neurodegenerative disease. Brain cell types can be differentiated from the iPSC's, cellular functions can be recapitulated using iPSC-derived cells in vitro and co-culture platforms can be used to know the complex interactions between brain cells during neurodegeneration. In the 4th and 5th decades of life, family AD (FAD) begins early and usually after the age of 70, sporadic later AD (SAD) occurs. Pathological features of AN – shared between FAD and SAD – are senile plates and neurofibrillary tangles (NFTs).[14]

The Genome-wide association studies (GWAS) identified numerous additional SAD risk genes that are expressed primarily in non-neuronal cells of the brain. Genome editing techniques can be used to identify the AD-linked mutations to examine phenotypes in isogenic backgrounds. Genetic Diversity can be a major hindrance in experimental analysis like while focusing on the function of the specific mutation, genetic changes may mask or exacerbate a phenotype.[14]CRISPR/Cas9 system can be used to examine the effects of targeted mutations in an otherwise identical (isogenic) genetic background.

The Astrocytes that are also involve in the progression of the decreasing, which took part in many functions in Central Nervous System(CNS). The degradation of these cells

may affect the normal physiological function, secreting the senescence-associated secretory phenotype (SASP) factors that causes A β accumulation, tau hyperphosphorylation, and the deposition of neurofibrillary tangles (NFTs) in Alzheimer's disease (AD).[15] These Astrocytes may cause the early onset of AD. Further studies are conducting to find the connection between cellular senescence and neurodegenerative diseases.

Preclinical trials are developed to find the causes for AD progression that might improve in the treatment, prevent cognitive decline of AD. Identification of depressive symptoms such as subclinical features and subsequent antidepressive treatment might help to halt the dementia. Individuals with subsyndromal depression including carriers of the apolipoprotein E (APOE) ϵ 4 allele, family history of AD or amyloid pathology with no or only subtle cognitive impairment, is crucial and could generate an opportunity for early intervention of dementia. More research is needed in the depression part as it is one of the risk factor for AD.[16]

Using GWAS studies the new genes were identifies for LOAD in which the GWAS data combined with high-quality imputation panels can reveal rare disease risk variants. Many rare variants were identified, yet many are need to be identify by using better imputation panels.[17] The rare variants will help in understanding the disease mechanisms and potential drug targets. Discovery of the risk genes at genome-wide loci remains challenging.

Transcriptome-wide meta-analysis was performed to encounter the Abnormal gene expression patterns may contribute to the onset and progression of late-onset Alzheimer's disease (LOAD). In this study, the top genes are implicated in pathways that are known to be important in AD pathophysiology such as synaptic processes (CREB5), endosomal sorting (FLOT1), immune system (CD46, IRAK3) and calcium homeostasis (TMBIM6). Expression profiling and eQTL studies provide a complementary approach to AD risk association in the identification of novel AD candidate genes and pathways.[18] Further studies are required to find the role of the genes that are identified.

A recent study found that the coat protein complex I (COPI)–dependent trafficking in vivo could significantly reduce amyloid plaques in the cortex and hippocampus of neurological in the AD mouse models and identified 12 single-nucleotide polymorphisms in COPI genes to be significantly associated with increased AD risk using 6,795 samples. The large-scale GWAS dataset to investigate the potential association between the COPI genes and AD susceptibility by both SNP and gene-based tests.[19] It is found that the GWAS results was match with one of the eQTLs result. uture studies are required to replicate these findings using large-scale GWAS and eQTLs datasets.

Using International Genomics of Alzheimer's Project (IGAP) many disease-associated genes have been identified but there is no enough information about how those disease-associated genes affect gene expression and disease-related pathways. By integrating GWAS summary data from IGAP and five different expression-level data by using the transcriptome-wide association study method disease-causal genes under strict multiple testing ($\alpha < 0.05$), and some genes are newly identified. Further research from areas such as metabolomics and proteomics is needed to study the disease association between AD and autoimmune diseases.[20]

In neurodegenerative diseases, miRNAs play a critical role in AD as well. MiRNAs are non-coding nucleic acids that regulate gene expression by binding to the complementary target messenger RNA (mRNA). In terms of neurite outgrowth, dendritic spine morphology, neuronal differentiation, and synaptic plasticity, miRNAs are involved.[21]The malfunction of miRNAs in AD has been found to be increasingly recognized and miRNAs have been named as biomarkers or therapeutic agents. MiRNAs theoretically involved in diagnostics will be used by miRNA panels for specific tasks such as disease level, risk prediction, and progression of the disease.

The TMEM106B gene rs1990622 variant is a risk factor for frontotemporal dementia (FTD). The rs1990622 T allele has been identified as the most common and occurs more frequently in females than in males. Using large-scale genome-wide association study (GWAS) datasets from the International Genomics of Alzheimer's Project and UK Biobank, methods such as genetic association analysis were conducted. Using the gene expression data in GTEx, a gene expression analysis of TMEM106B in 49

different human tissues was performed.[16] Multiple datasets from UKBEC, GTEx, and the Mayo RNAseq Study were used to analyse expression quantitative trait loci (eQTLs).

In order to provide evidence of the pair of ADS GWAS and eQTLs affecting both AD and the TMEM106B expression in a specific region, colocalization analysis was performed. Analysis of gene expression showed that TMEM106B had high expression in the cerebellar hemisphere, tibial nerve, cerebellum, and spinal cord, but low expression in 10 other human brain tissues, including the frontal cortex, hypothalamus, nucleus accumbens, caudate, substantia nigra, cortex, hippocampus, amygdala, and putamen. The disadvantage in this research is that the data sets used in this research are not accessible to the public, so more research will be done when available.[22]

Human AD-specific transcriptomic alterations conserved across multiple cohorts and cortical regions were discovered using a functional genomics and integrative systems biology approach. A meta-analytical approach was carried out on more than 1200 postmortem brain RNA-seq samples across brain banks and identified robust disease-related co-expression modules for consensus weighted gene co-expression network analysis (cWGCNA). Using this approach, it was identified the brain-region specific alterations that are fundamental to disease pathophysiology.

To define transcription factors (TFs) regulating the neuronal and glial modules, generic and brain-specific ATAC-seq and ChIP-seq datasets were used, identifying members of the NF- κ B pathway as regulators of changes in AD-associated co-expression. To provide a comprehensive picture of the genetic alterations associated with AD, DNA methylation, histone acetylation (H3K9ac) and expression and methylation quantitative loci (eQTLs and mQTLs) were used.[23]

Via genome-wide association studies (GWAS) and next-generation sequencing studies on increasingly broad AD populations, the field of AD genetics is still aggressively exploring additional genetic risk loci, a trend emerging to simultaneously interrogate transcriptomic data to research the impact of newly established genetic risk factors at the transcriptome stage. [9] Transcriptome enhancement for further studies on the

functional pathways involved in AD by the use of various sequencing technologies such as second and third generation sequencing techniques, single-cell, bioinformatic research. [24]

In AD, genome and transcriptome play an important role primarily in large amounts of research on genome wide association studies (GWAS) and profiling of gene expression. By considering the entire genome of the disease, GWAS played a major part in defining the SNPs or genetic variants, but the role of those variants remains uncertain. It is assumed in transcriptomic studies that a disease transcriptome would give the DEG involved in the progression of the disease, but in fact, there are several false positives that not all DEGs are involved in the progression of the disease. This can be solved by combining the genome and transcriptome of AD and by decreasing the false positives to increase the confidence level of the outcome. In order to cross-validate the findings, more analysis can be performed through Gene enrichment studies and Pathway inference studies.

CHAPTER 3

AIM AND OBJECTIVE

The aim of the study is to identify differentially expressed genes (DEGs) for the analysis of novel genes and false positives in Alzheimer's disease by integrating two Omics studies.

Listed are the prominent objectives of the study:

- ✓ To find the SNPs in the genome of the whole genome of single-cell brain cells of patients with Alzheimer's disease.
- ✓ To integrate two different omic studies to reduce the false positive in AD.
- ✓ To find up-regulated & down-regulated genes between AD and non-AD samples.
- ✓ To find the probabilities of the SNPs those are present in the genome of the AD sample.
- ✓ To identify the DEG novel genes present in the transcriptome of AD.
- ✓ To identify the population frequency of each SNP.
- ✓ To identify the SNIPS present in genotypes and phenotypes in AD patients.

CHAPTER 4

MATERIALS AND METHODS

4.1 MATERIALS THAT ARE USED

4.1.1 Sample Collection

4.1.1.1 NCBI

National Centre for Biotechnology Information (NCBI). This is a part of United States Library of Medicine (NLM) and a branch of National Institutes of Health (NIH) funded and approved by government of United States. NCBI located in Bethesda, Maryland founded in 1988 by Senator Claude Pepper.

NCBI contains series of databases related to biotechnology and biomedicine and it is a major resource for bioinformatics tools and services.

It has many databases such as Nucleotide, PubMed Central, PubMed, SNP, Gene, Protein, PubChem, SNP and many more.

It also provides with wide number of free open resources such as DNA & RNA, Homology, Genomes & Maps, Sequence analysis, Variation, Chemicals & Bioassays, Taxonomy, Literature, Homology etc.

The database can be selected based on the study we are focusing on.

Each database will have its own significance and functions with detailed information.

4.1.1.2 SRA

Sequence Read Archive (SRA) is a database in NCBI. It is a freely available public repository for storing large amount of information such as High-through put sequencing data and also meta-analysis data.

It takes information from wide number of resources such as Meta genomics and environment samples.

SRA mainly contains the raw sequence such as reads and sequence alignment information which can be used in different areas of Data analysis.

Most of the NGS data will be incorporated in the SRA with the sample number and the total number of runs in the sequence and other related information.

In this study the sample datasets that are taken were Single-cell whole genome sequencing of brain cells of Alzheimer's disease patients of SRA accession number ERX1461867 for GWAs studies and the sample dataset Small RNA sequencing of brain-derived extracellular vesicles from post-mortem brain tissue of controls and Alzheimer's disease patients with different APOE genotypes of SRA accession number SRX9296596, SRX9296593, SRX9296590.

- Sample 1 was collected from a female patient of age 64. The part of the sample was isolated from the frontal cortex individuals with mild AD (Braak stage III) and with advanced AD (Braak stage VI). Single-cell whole genome sequencing (scWGS) was used to assess aneuploidy. Sequencing was performed under Illumina platform by using Illumina Hiseq 2500 instrument with library source as genome and with single library layout.

The screenshot displays the SRA Run Selector web application. At the top, the NCBI logo and 'SRA Run Selector' title are visible. A search bar contains the accession number 'ERP015306'. On the left, a 'Filters List' shows various criteria like AGE, Bases, Bytes, Cell_type, Description, disease_staging, Individual, INSDC_first_public, and sex, each with a checkbox. The 'Common Fields' section provides detailed metadata for the selected accession, including BioProject (PRJEB13739), Consent (PUBLIC), Assay Type (WGS), AvgSpotLen (51), Broker_name (ArrayExpress), Center Name (Nancy Halsema, Inge Kazemier), DATASTORE filetype (FASTQ, SRA), DATASTORE provider (ENA, GS, NCBI, S3), and DATASTORE region (ena, gs, US, ncbi, public, s3.us-east-1). Below this, a 'Select' table shows the total number of runs (1027), bytes (17.85 Gb), and bases (43.74 Gb). It also includes options for downloading metadata or an accession list, and buttons for 'Cloud Data Delivery' and 'Computing'.

Fig:4.1 : Sample of genome dataset collection in SRA Run selector

- Sample 2 (3 datasets) was isolated from the post-mortem brain tissues by separating the brain-derived EVs (bdEVs) in 23 AD patients with different APOE genotypes and 7 controls. Sequencing was performed under ION TORRENT by using Ion Torrent S5 instrument with library source as transcriptome and with single library layout.

The screenshot shows the NCBI SRA Run Selector interface. At the top, the NCBI logo and 'SRA Run Selector' are visible. A search bar contains the accession 'SRP287832'. On the left, a 'Filters List' shows various filters like 'apoe_genotype', 'AvgSpotLen', 'Bases', 'Bytes', 'disease_state', and 'Treatment'. The 'Common Fields' section displays metadata for the selected accession: BioProject (PRJNA670004), Consent (PUBLIC), Assay Type (RNA-Seq), Center Name (GEO), DATASTORE filetype (BAM, SRA), DATASTORE provider (GS, NCBI, S3), DATASTORE region (gs.US, ncbi.public, s3.us-east-1), Instrument (Ion Torrent S5), and LibraryLayout (SINGLE). Below this, a table summarizes the data:

Select	Runs	Bytes	Bases	Download	Cloud Data Delivery	Computing
Total	90	5.08 Gb	6.63 G	Metadata or Accession List		
Selected	0	0	0	Metadata or Accession List or JWT Cart	Deliver Data	Galaxy

Fig:4.2 : Sample of Transcriptome data collection in SRA Run Selector

4.1.2 Faster download and extract read in FASTA format from NCBI SRA

This will give four different out puts like single end data, paired end data, other data and fasterq dump log. All the files are dumped files which are uploaded from a single SRA data .It will tell us which type of data we have like single end data , paired end data , other data and fasterq dump log.

The data from the SRA will be downloaded in fastq format. This will aid in the fast alignment of the series and provide results for the provided input.

4.1.3 FastQC Read Quality Report

FastQC offers an easy way to conduct some quality control tests on raw sequence data from pipelines with high throughput sequencing. This gives a modular collection of analyses that can be used to provide a simple indication of whether the knowledge has any concerns that you should be mindful of before doing any further research. The quality reports will give the understanding about how much the reads are accurate with basic statistics of the raw sequence.

It takes the input as FASTQ files ith a Quality check protocol and gives the human readable report.[25] The quality check will be done by considering the position of the reads, adaptor content and many other parameters.

4.1.4 Trimmomatic

For Illumina paired-end and single-end reads, Trimmomatic conducts a number of valuable quality control tasks. As a sequence of steps, these activities are carried out. It is necessary to define at least one step, and steps are run in the following order. Trimmomatic is used to trim the reads and other adaptor sequences and it will crop the FASTQC Illumina data. Trimming performs on Paired end data and single end data. There are many sequence of steps that takes place during trimming such as Illumina clip, Sliding window, Maxinfo, Leading, Trailing, Crop, Headcrop, Minlen, Avgqual, Tophred33, Tophred64.

4.1.5 HISAT2

For mapping RNA-sequence reads, HISAT2 is a fast and sensitive spliced alignment programme. HISAT2 uses a large number of small FM indexes to cover the entire genome, in addition to one global FM index that covers the entire genome. Each index represents a genomic area of 64,000 bp, and 48,000 indexes are required to cover the entire human genome. HISAT2 was developed by implementing HISAT and Bowtie. The output will be in the SAM format for HISAT2.

4.1.6 Cufflink

In RNA-Sequence samples, Cufflinks assembles transcripts, estimates their abundances, and searches for differential expression and regulation. It takes aligned RNA-Sequence reads and condenses them into a manageable set of transcripts. Cufflinks then measures the relative abundances of these transcripts based on the number of reads that endorse each one, accounting for library preparation biases. It runs on intel based computers with linux and Mac OS X.

4.1.7 Cuffmerge

Cuffmerge to merge several Cufflinks assemblies together. It also runs Cuffcompare for you and filters out a variety of transcriptomes that are most likely objects. If you have a reference GTF file, you can use it to help the script merge novel and known isoforms gracefully and boost overall assembly efficiency.

It combines the information obtained from the cufflink, and the results will be saved in a GTf file format. It includes the chromosome number, and tells whether an exon or intron is present. It will also include the gene id, exon number, and all information relevant to that gene.

4.1.8 Cuffdiff

As input, the Cuffdiff module accepts two or more fragment alignment SAM/BAM files from TopHat or another read aligner, as well as a GTF file containing transcript annotations such as merged Gtf from Cuffmerge. Cuffdiff creates a variety of output files, each of which includes test results for changes in transcript, primary transcript, and gene expression.

4.1.9 Bowtie2

Bowtie2 is a memory-efficient short read aligner designed to align large sets of short DNA sequences (reads) to large genomes easily. On a standard workstation, it aligns 35-base-pair reads to the human genome at a rate of 25 million reads per hour. Bowtie is also the base for other instruments, such as the TopHat RNA-sequence reads fast splice junction mapper. Bowtie2 supports different alignment modes such as local, paired end and gapped. Bowtie2 can be used in constructing pipelines for variant calling, RNA-seq, CHIP-seq in comparative genomics. Advantage of bowtie2 is it doesn't require to align end-to-end. It supports local alignment on which the reads are trimmed at both the ends. Bowtie2 doesn't has upper limit on the read length.[26,27] There are different packages such as Bioconda for Bowtie2. Bowtie2 gives the results in SAM format.

4.1.10 SAM

Heng Li and Bob Handsaker et al. created the Sequence Alignment Map (SAM), a text-based format for storing biological sequences aligned to a reference sequence. It's commonly used for storing data produced by next-generation sequencing technologies, such as nucleotide sequences, and the standard has been extended to include unmapped sequences. SAM is a flexible and free open software platform for parsing

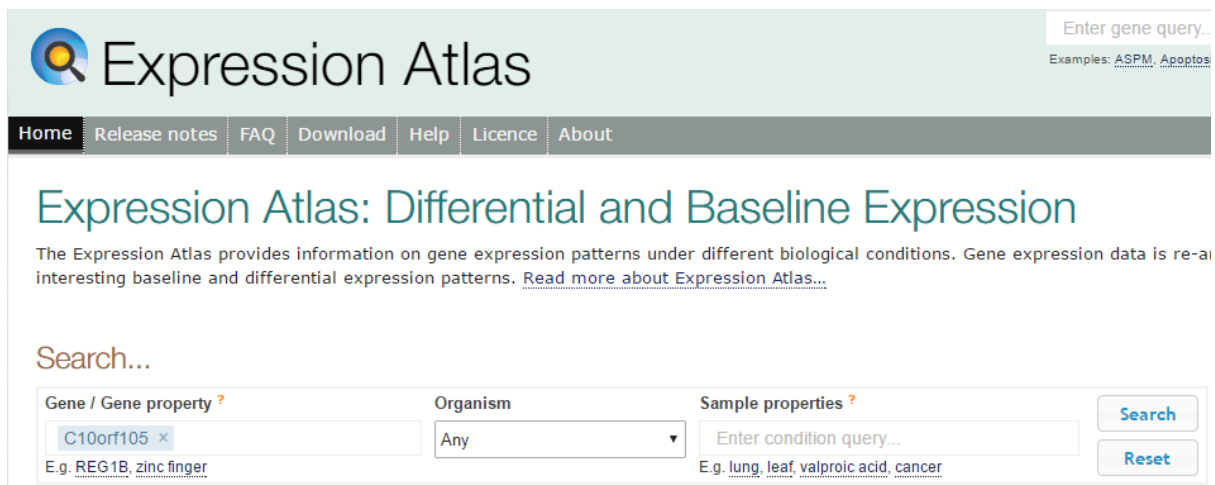
and manipulating the alignments in SAM/BAM format with different sequencing platforms.[28] It takes post-processing alignments in SAM format like variant caller, alignment viewer. SAM involves in removing the duplicates, alignment, generate per-position in pileup format.

4.1.11 Variant calling

The method of distinguishing variants from sequence data is known as variant calling. It will filter out all the duplicate reads. The output will be in VCF format for storing the variation data. Millions of variants can be stored in a single VCF file. Variant callers take in the BAM file and return a list of variants(SNP).[29] It is done to find the difference between the sequence of the patient sample with the reference sequence. Differences such as SNPs, INDELS, CNVs, SVs.

4.1.12 Expression Atlas

The European Bioinformatics Institute maintains the Expression Atlas database, which contains information on gene expression patterns from RNA-Sequence and Microarray studies, as well as protein expression from proteomics studies.



The screenshot shows the Expression Atlas website. At the top, there is a search bar labeled "Enter gene query.." with examples "ASPM, Apoptosis". Below this is a navigation bar with links: Home, Release notes, FAQ, Download, Help, Licence, About. The main heading is "Expression Atlas: Differential and Baseline Expression". Below the heading, a paragraph states: "The Expression Atlas provides information on gene expression patterns under different biological conditions. Gene expression data is re-analyzed to identify interesting baseline and differential expression patterns. [Read more about Expression Atlas...](#)".

Below the paragraph is a "Search..." section with three input fields and two buttons. The first field is labeled "Gene / Gene property ?" and contains "C10orf105" with a dropdown arrow and a link "E.g. REG1B, zinc finger". The second field is labeled "Organism" and contains "Any" with a dropdown arrow. The third field is labeled "Sample properties ?" and contains "Enter condition query..." with a link "E.g. lung, leaf, valproic acid, cancer". There are "Search" and "Reset" buttons to the right of the input fields.

Fig:4.3 : Expression Atlas for DEG

There are three check boxes on it. The first search box is for gene gene properties, and the second search box is for species, which includes all types of organisms. The third search box is for sample properties, which includes all body organ samples.

4.1.13 PhenGenI

The Phenotype-Genotype Integrator (PheGenI) combines data from NHGRI's genome-wide association analysis (GWAS) catalogue with information from multiple NCBI databases, including Gene, dbGaP, OMIM, eQTL, and dbSNP. Follow up on GWAS findings, will help with prioritisation. This phenotype-focused resource, designed for clinicians and epidemiologists who want this will include the organism's phenotypic and genotypic characteristics.

The phenotypic and genotypic search boxes each have three different items in them, such as locations, genes, and snp chromosome and range

4.1.14 GO Annotation

A Gene Ontology (GO) annotation is a statement about a gene's role. By associating a gene or gene product with a GO word, GO annotations are generated. These statements together make up a snapshot of existing biological understanding. Gene-based product. Statements about how a gene works at the molecular level, where it functions in the cell, and what biological processes (pathways, programmes) it aids in are all captured in GO annotations. A GO annotation is uniquely identified by four pieces of information. Although there are additional components, such as qualifiers and annotations, that a curator may use to indicate further details.

4.1.15 Genome Asia 100K

This includes a reference whole-genome sequencing dataset from 1,739 people from 219 ethnic groups and 64 countries across Asia. Genetic variation, population composition, disease interactions, and founder effects are all catalogued.[30] We also look at using this dataset in imputation to help with genetic studies in Asian and global populations.

4.1.16 IGAP

The summary results data from the 2013 meta-analysis of Genome-wide Association (GWA) data in Alzheimer's disease has been released by the International Genomics of

Alzheimer's Project (IGAP), allowing other researchers to look for evidence of association with specific variants or loci.

4.1.17 Manhattan Plot

A Manhattan plot, which plots the statistical significance of an association as $-\log_{10}(\text{p-value})$ on the y-axis against chromosomes on the x-axis, is a good way to show millions of genetic variants in one figure. Regions of the genome that cross a certain significance threshold are easily identifiable. is a type of scatter plot that is most commonly used to display data with a large number of data points, many of which have non-zero amplitude, and a distribution of higher-magnitude values. The plot is commonly used to display significant SNPs in genome-wide association studies (GWAS).

4.1.18 Volcano Plot

A volcano plot is a type of scatter plot that depicts statistical significance (P value) versus change magnitude (fold change). It allows for the rapid visual identification of genes with large fold changes that are statistically significant. These genes could be the most important in terms of biology. The negative log of the p value is plotted on the y axis to create a volcano plot (usually base 10). Data points with low p values (highly significant) appear near the top of the plot as a result of this. The log of the fold change between the two conditions is plotted on the x axis.

4.1.19 DEG HeatMap

The heatmap can also be used in conjunction with clustering methods, which group genes and/or samples together based on their gene expression patterns' similarity. Data is displayed in a grid in heat maps, with each row representing a gene and each column representing a sample. A heatmap is a graphical representation of data that employs a color-coding system to represent various values. Heatmaps are used in a variety of analytics applications, but they're most commonly used to display user behaviour on specific webpages or webpage templates.

4.2 METHODOLOGY

4.2.1 Samples

The accession number was used to collect data from SRA in NCBI. ERX1461867, SRX9296596, SRX9296593, and SRX9296590 are the accession numbers for each data set. The raw sequence of the samples generated by sequencing techniques is contained in this data. The total number of runs for each dataset is chosen using the SRA Run selector. The sequence will be in FASTQ format if you download the accession list.

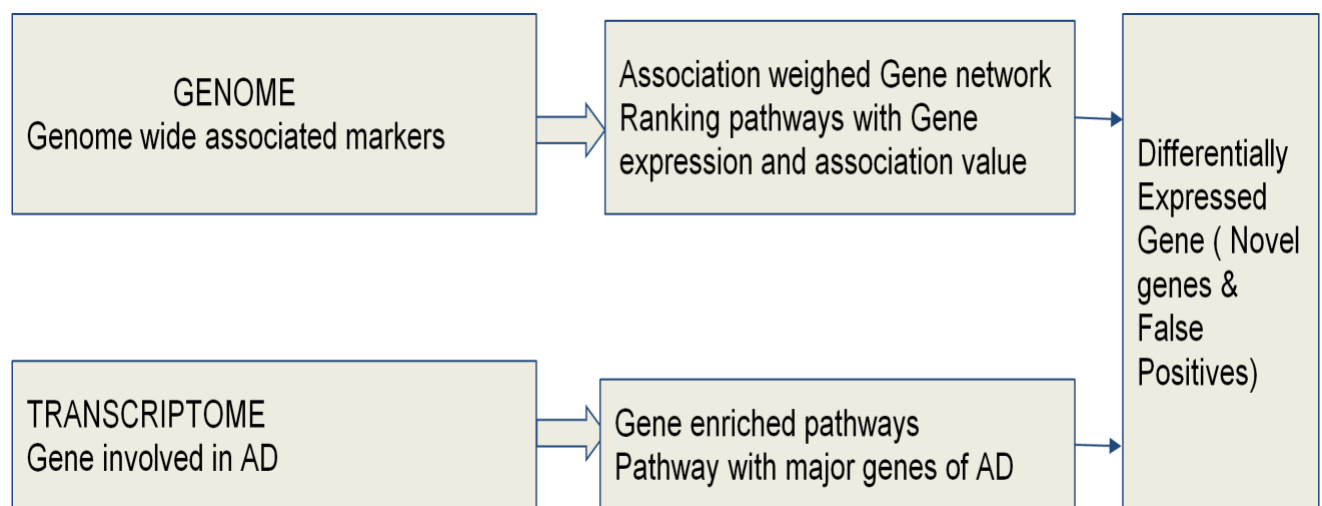


Fig:4.4 : Overall Workflow in finding the Differentially Expressed Genes

4.2.2 Extract reads from FASTQ (Primary analysis)

Faster download and extract reads in fastq format should be given to the Download FASTQ list; it will extract the reads from the FASTQ into different types of data such as single end data, paired end data, other data, and fasterq dump logs. We will identify the data of the provided input data using that data. If the data is single end, there is only one end; if the data is paired end, there are two ends; and the output can be any of these different types of data depending on the sequence we used.

4.2.3 Quality check

The single end data was used as input for the quality check. It will run a quality control test on raw sequence data before presenting a modular series of analyses with simple indicators. It will display a diagrammatic representation of the sequence parameters, as well as a summary of the sequence parameters, including failure name, encoding type, total sequence duration, and GC percentage. Basic Statistics, Per base sequence quality, Per sequence quality scores, Per base sequence content, Per base GC content, Per sequence GC content, Per base N content, Sequence Length Distribution, Sequence Duplication Levels, Overrepresented sequences, Kmer Content will be included in the HTML output.

4.2.4 Trimming

Trimmomatic is a multithreaded command line tool for trimming and cropping Illumina (FASTQ) data and deleting adapters. Trimmomatic receives the raw sequence reads and trims out any unwanted reads or adapters. It can be used for both paired and single-end reads. It completes the task in a step-by-step manner in order to eliminate short reads and reduce the overall number of read sequences. Adapter trimming removes adapter sequences from the 3' ends of reads in the Illumina FASTQ file generation pipelines.

Because adapter sequences in reads obstruct downstream analyses like read alignment to a guide, they should be removed. Trim one or both ends of the reads, leaving only a region of high-quality bases... Setting a threshold value for low-quality base calls accomplishes this (using the -c option). The default value is 20, indicating that quality scores of less than 20 are poor. The trimmomatic output is forwarded to the next stage of the procedure.

4.2.5 Alignment (secondary analysis)

4.2.5.1 Bowtie2

Burrows-Wheeler (Burrows-Wheeler) For alignment of raw sequences with reference sequences, Aligners or Bowtie can be used. The trimmed raw sequence was used as

input for bowtie alignment with the reference sequence Human (Homo sapiens): hg38. It will give the results based on the mapping status after the alignment is completed. The overall alignment rate was calculated using the number of unpaired reads, aligned reads, number of reads that are aligned exactly one time, and number of reads that are aligned more than one time. The final product will be a SAM file.

4.2.5.2 HISAT2

HISAT2 is a quick and responsive alignment software for mapping next-generation sequencing reads (both DNA and RNA) to a population of human genomes as well as a single reference genome for mapping RNA sequence reads. The alignment is done between the raw sequence which was trimmed and the reference Human (Homo sapiens): Hg38. HISAT2 employs a large number of small GFM indexes to cover the entire genome. These small indexes, when combined with a variety of alignment techniques, allow for rapid and precise alignment of sequencing reads. A Hierarchical Graph FM index is the name of this new indexing system (HGFM). Python 3 is supported. The output format will be SAM which will include total summery aligned 0 times, 1 time, >1 time, and overall alignment rate.

The performance of HISAT2 is used in more cufflink research.

4.2.6 SAM tools

For the filtering and results in BAM format, the bowtie alignment was given in SAMtools with several parameters such as

QName: Query template name

Flag: Combination of bitwise FLAGS

RName: Reference sequence name of alignment

POS: Position of first CIGAR operation that consumes a reference base.

MAPQ: Mapping Quality

CIGAR: CIGAR string

RNEXT: Reference sequence name of the primary alignment

TLEN: Signed observed the NEXT read in template

PNEXT: Position of the primary alignment of NEXT read in the template

SEQ: Segment sequence

QUAL: ASCII of base quality plus 33

4.2.7 Variant Calling

In the variant calling, the BAM file generated by SAMtools was used to check the base quality score, mapping quality, and base alignment quality. The VCF file with the list of variants is the variant calling output. The millions of variants are stored in the VCF file (SNIPS). Chromosome number, position, reference nucleotide, altered nucleotide, Quality, filter, and Info fields are all included in the file format. The raw depth, allele frequency, Phred-scaled strand bias, and counts for ref-forward, ref-reverse, alt-forward, and alt-reverse bases are all contained in the INFO field.

4.2.8 SNP Annotation

It's used to annotate IDs and to extract data from dbSNP. The SNPsift annotator was given the dbSNP file, and the resultant format was VCF with the SNP rs IDs (Reference SNP report). A single nucleotide or a single portion of a DNA sequence can have a variation or mutation called an SNP. It can occur in encoded aminoacids (nonsynonymous), silent (synonymous), or non-coding regions. Individual diversity, genome evolution, family traits such as curly hair, inter individual differences in drug response, diseases such as diabetes, obesity, hypertension, psychiatric disorders, and so on may all be affected by SNIPS. SNPs can affect or trigger promoter activity (gene expression), mRNA, protein stability, and protein expression, all of which can lead to disease. Identifying different gene variations (SNIPS) and analysing their effects can help us understand their impact on gene function and individual health.

4.2.9 Transcript assembly

4.2.9.1 Cufflinks

Cufflinks assembles transcripts, calculates their abundances, and looks for differential expression and regulation in RNA-Sequence samples. It condenses synchronised RNA-Sequence reads into a manageable number of transcripts. Cufflinks then uses the number of reads to calculate the relative abundances of these transcripts. Cufflinks is a collection of methods for analysing RNA-sequence data for differential gene expression. It gathers matched reads from a series of transcripts and calculates relative abundances. The output will be in the gtf file format. Combine multiple Cufflinks assemblies into one. It also takes care of running Cuffcompare for you and filters a variety of transfrags that are most likely artefacts. If you have a reference GTF file, you can give it to the script to help it merge novel and known isoforms gracefully and improve overall assembly efficiency. The primary goal of this script is to create as input, cuffmerge accepts multiple assembly GTF files from Cufflinks. The complete paths to the input GTF files must be defined in a "manifest" file. You want to combine multiple GTF files into a single GTF file. The input assemblies are combined with the reference GTF to create the final product.

The data from the cufflink will be combined, and the results will be stored in the GTf file format. The chromosome number, as well as whether an exon or intron is present, will be included. It will also contain the gene's id, exon number, and any other relevant information.

4.2.5.2 Cuffmerge

Cuffmerge output will be used in the next stage, cuffdiff. The Cuffdiff module accepts two or more fragment alignment SAM/BAM files from TopHat or another read aligner, as well as a GTF file containing transcript annotations from Cuffmerge, such as merged. Gtf. Cuffdiff generates a number of output files, each of which contains test results for transcript changes, as well as the primary transcript. Includes monitoring id class code closest reference id gene id short gene name length and FPKM differential expression checking. As well as gene expression.

4.2.5.3 Cuffdiff

The Cuffdiff generates differential expression outputs. The remaining outputs determine the likelihood of a substantial shift in the balance of any locus between any pairwise measures. It's worth noting that such files only contain findings for genes with at least two splice variants. Comparing gene and transcript expression levels in RNA-Sequence experiments is a difficult task. Cuffdiff is a powerful method for making these comparisons, revealing not only which genes are up- or down-regulated between two or more conditions, but also which genes are differentially spliced or undergoing other forms of isoform-level control.

4.2.10 Identifying SNPs in genotype and phenotype

The source for the Phenotype traits of Alzheimer's Disease was PhenGenI (Phenotype-Genotype Integrator). PhenGenI will provide information on association results, genes, SNPS, eQTL data, dbGap studies, and genome view. A list of rs ids, gene, location, and P-value are included in the association results. The meta analysis genome data of Alzheimer's disease was taken using IGAP (International Genomics of Alzheimer's Project). The stage 1 results were taken from the data file that was downloaded. Compare and extract the common rs in d's both PhenGenI and IGAP to get the common SNPs that are found in both genotype and phenotype.

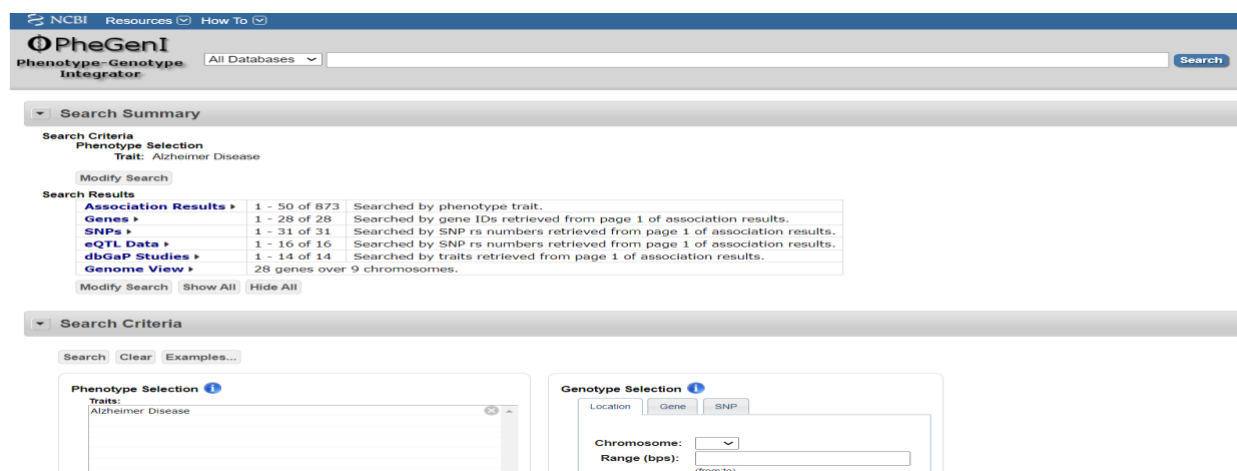


Fig:4.5 :PhenGenI summary results for Alzheimer's disease

Association Results											
1 - 50 of 873 < Previous Next > Page 1 Go Download Modify Search											
#	Trait	rs #	Context	Gene	Location	P-value	Source	Study	Population	PubMed	
1	Alzheimer Disease	rs2075650	Intron	TOMM40	19: 44,892,362	1.000 x 10 ⁻²⁸⁵	NHGRI		European Hispanic	20460622	
2	Alzheimer Disease	rs2075650	Intron	TOMM40	19: 44,892,362	2.000 x 10 ⁻¹⁵⁷	NHGRI		European	19734902	
3	Alzheimer Disease	rs4420638	nearGene-3	APOC1	19: 44,919,689	8.000 x 10 ⁻¹⁴⁹	NHGRI		European	22832961	
4	Alzheimer Disease	rs2075650	Intron	TOMM40	19: 44,892,362	9.000 x 10 ⁻¹¹⁶	NHGRI		European	24755620	
5	Alzheimer Disease	rs769449	Intron	APOE	19: 44,906,745	5.000 x 10 ⁻²⁴	NHGRI		NR	28247064	
6	Alzheimer Disease	rs157580	Intron	TOMM40	19: 44,892,009	8.000 x 10 ⁻⁸⁹	NHGRI		European	21627779	
7	Alzheimer Disease	rs61812598	Intron	IL6R	1: 154,447,611	6.000 x 10 ⁻⁶³	NHGRI		NR	25340798	
8	Alzheimer Disease	rs6857	UTR-3	PVRL2	19: 44,888,997	2.000 x 10 ⁻⁶²	NHGRI		NR	25188341	
9	Alzheimer Disease	rs157582	Intron	TOMM40	19: 44,892,962	9.000 x 10 ⁻⁵²	NHGRI		African European African American Indian American	22005930	
10	Alzheimer Disease	rs573521	Intergenic	MMP3, MMP12	11: 102,846,249	2.000 x 10 ⁻⁴⁴	NHGRI		NR	25340798	
11	Alzheimer Disease	rs4420638	nearGene-3	APOC1	19: 44,919,689	2.000 x 10 ⁻⁴⁴	NHGRI		European	17998437	
12	Alzheimer Disease	rs6733839	Intergenic	BIN1, CYP27C1	2: 127,135,234	7.000 x 10 ⁻⁴⁴	NHGRI		European	24162737	
13	Alzheimer Disease	rs157580	Intron	TOMM40	19: 44,892,009	1.000 x 10 ⁻⁴²	NHGRI		European	19125160	
14	Alzheimer Disease	rs4420638	nearGene-3	APOC1	19: 44,919,689	1.000 x 10 ⁻³⁹	NHGRI		European	17474819	
15	Alzheimer Disease	rs4420638	nearGene-3	APOC1	19: 44,919,689	1.000 x 10 ⁻³⁸	NHGRI		NR	17975299	
16	Alzheimer Disease	rs519113	Intron	PVRL2	19: 44,873,027	5.000 x 10 ⁻³⁸	NHGRI		East Asian Asian European	23565137	
17	Alzheimer Disease	rs6857	UTR-3	PVRL2	19: 44,888,997	3.000 x 10 ⁻³⁸	NHGRI		NR	25188341	
18	Alzheimer Disease	rs2075650	Intron	TOMM40	19: 44,892,362	5.000 x 10 ⁻³⁸	NHGRI		NR	20885792	
19	Alzheimer Disease	rs769449	Intron	APOE	19: 44,906,745	5.000 x 10 ⁻³³	NHGRI		NR	28247064	
20	Alzheimer Disease	rs769449	Intron	APOE	19: 44,906,745	4.000 x 10 ⁻²⁹	NHGRI		NR	28247064	
21	Alzheimer Disease	rs10792832	Intergenic	PICALM, FNTA1	11: 86,156,833	9.000 x 10 ⁻²⁸	NHGRI		European	24162737	
22	Alzheimer Disease	rs9331896	Intron	CLU	8: 27,610,169	3.000 x 10 ⁻²⁸	NHGRI		European	24162737	
23	Alzheimer Disease	rs9331896	nearGene-3	MIR6843	8: 27,610,169	3.000 x 10 ⁻²⁵	NHGRI		European	24162737	
24	Alzheimer Disease	rs56131196	nearGene-3	APOC1	19: 44,919,589	3.000 x 10 ⁻²⁴	NHGRI		European	26830138	
25	Alzheimer Disease	rs6656401	Intron	CR1	1: 207,518,704	6.000 x 10 ⁻²⁴	NHGRI		European	24162737	

Fig:4.6 :PhenGenI Association table for Alzheimer's disease

C1 MarkerName																	
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Chromosome	Position	MarkerName	Effect_allele	Non_Effect_allele	Beta	SE	Pvalue									
2	1	751343	rs28544273	A	T	-0.0146	0.0338	0.6651									
3	1	751756	rs143225517	C	T	-0.0146	0.0338	0.6651									
4	1	752566	rs3094315	G	A	-0.0122	0.0294	0.6773									
5	1	753405	rs61770173	C	A	-0.0126	0.0339	0.7104									
6	1	768253	rs2977608	A	C	-0.0394	0.0261	0.1308									
7	1	768448	rs77786510	A	G	-0.0385	0.0303	0.2034									
8	1	769963	rs7518545	A	G	-0.0471	0.036	0.1904									
9	1	845274	rs112856858	T	G	0.0234	0.0329	0.4766									
10	1	845635	rs117086422	T	C	0.0317	0.0303	0.2944									
11	1	845938	rs57760052	A	G	0.0307	0.0295	0.2967									
12	1	846078	rs28612348	T	C	0.0289	0.0292	0.3224									
13	1	846398	rs58781670	A	G	0.0273	0.0298	0.3597									
14	1	846864	rs950122	C	G	0.0293	0.0296	0.322									
15	1	847228	rs3905286	T	C	0.0286	0.0297	0.336									
16	1	847491	rs28407778	A	G	0.0296	0.0301	0.326									
17	1	848090	rs4246505	A	G	0.0292	0.0308	0.3418									
18	1	848445	rs4626817	A	G	0.0288	0.0307	0.3479									
19	1	848456	rs11507767	G	A	0.0288	0.0307	0.3484									
20	1	848738	rs3829741	T	C	0.0292	0.0306	0.3397									
21	1	850062	rs28723578	T	A	0.0288	0.0307	0.3483									
22	1	850123	rs28622257	T	C	0.0282	0.0308	0.3602									
23	1	851190	rs28609852	A	G	0.0219	0.0308	0.4762									
24	1	851204	rs28552953	C	G	0.0267	0.0314	0.3942									
25	1	852664	rs28605311	T	C	0.023	0.0317	0.4694									
26	1	852758	rs4970462	C	G	0.0229	0.0317	0.4702									
27	1	935222	rs2298214	C	A	-0.042	0.0295	0.1536									
28	1	936111	rs1936360	C	T	-0.0348	0.0284	0.2197									
29	1	940005	rs2799056	G	A	-0.0293	0.0274	0.2852									
30	1	940006	rs4503704	C	T	0.0361	0.0367	0.176									

Fig:4.7 : IGAP meta-analysis data of GWAS data in Alzheimer's disease

4.2.11 Population Genetics

The SNIPS discovered through variant calling are included in the Genome Asia 100k in order to determine the SNP distribution among the population. The allele frequency of each SNP is calculated in this way.

4.2.12 GWAS data

The SNIPS found through variant calling are compared to the IGAP data, which includes a meta-analysis of AD genome-wide association data. It contains the P-value, Beta, and SE values, which are compared and SNIPS that are common are extracted.

CHAPTER 5

RESULTS AND DISCUSSION

5.1 QUALITY CHECK REPORT OF THE RAW SEQUENCES

The first step is to examine the raw sequence. The raw sequence was checked for quality to ensure that the data obtained from high-throughput techniques such as next-generation sequencing was accurate (Illumina, Ion Torrent).

Results of Quality check consists of

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication levels
- Overrepresented sequences
- Adapter Content

5.1.1 Base statistics of raw sequence

-ERR1461867, SRX9296596, SRX9296593, SRX9296590

Measure	Value
Filename	ERR1378585.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	884170
Sequences flagged as poor quality	0
Sequence length	51
%GC	42

Fig:5.1 : Base statistics of the raw sequence of run ERR1378585

Measure	Value
Filename	SRR12828917.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1870668
Sequences flagged as poor quality	0
Sequence length	8-340
%GC	52

Fig:5.2 : Base statistics of the raw sequence of run SRR12828917

The base statistics will include information such as the sequencing method, total number of sequences in each run, sequence length, and GC percentage for the overall raw sequence. Each run in the raw sequence will have different statistics.

5.1.2 Sequence quality check

-ERX1461867, SRX9296596, SRX9296593, SRX9296590

The sequence will be considered good in the base quality check based on the phred score. The base pairs are considered to be very accurate if they fall in the green zone, i.e. the phred score is above 28 or 30. It is considered less accurate if the base pairs in the yellow region, i.e. phred core, fall below 28 or above 20. If the base pairs in the red zone fall below 20, they are not considered to be of good quality. If the phred score is 20 or lower, there is a 1 in 100 chance of error. If the phred score is 20 or higher, there is a 1 in 1000 chance of error. All of the remaining runs were completed using the same method.

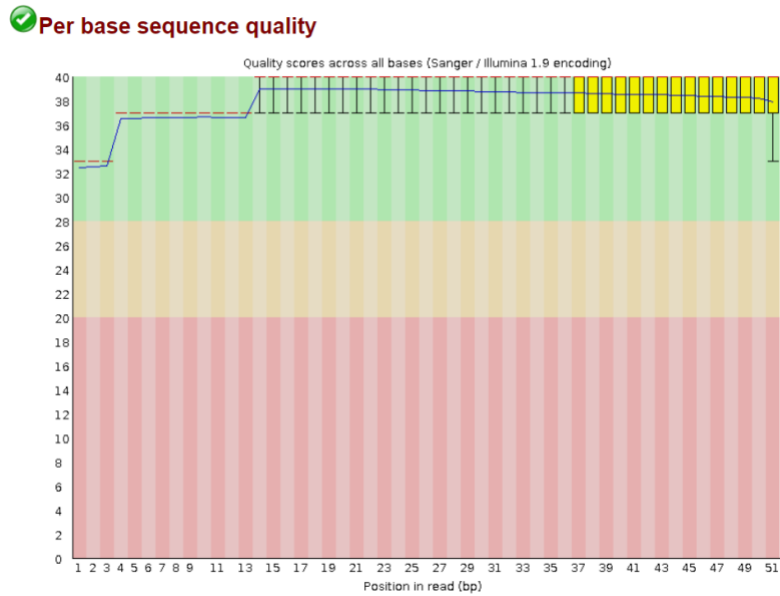


Fig:5.3 : Per base sequence quality report of raw sequence of run ERR1378585

The base quality is checked, and it is extremely accurate, with a phred score of >30.

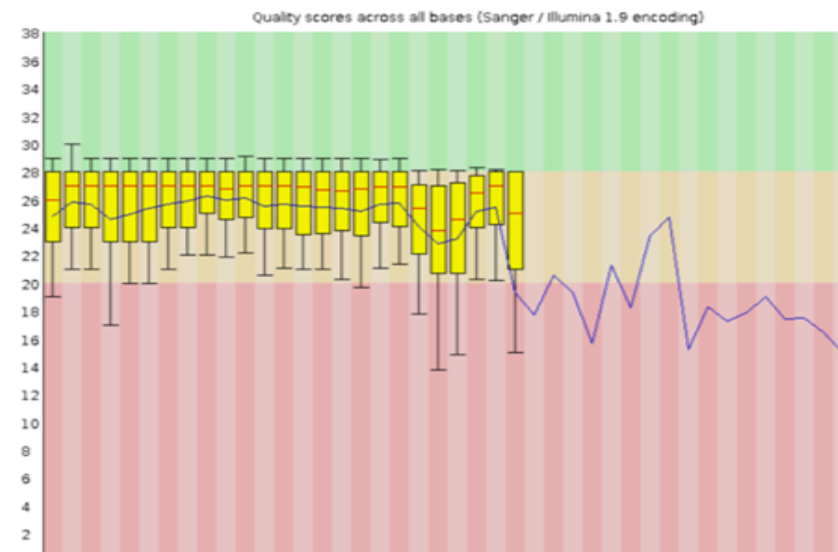


Fig:5.4 : Per base sequence quality report of raw sequence of run SRR12828917

With a phred score of 20 to >28, the base quality is checked, and the quality is less accurate.

5.1.3 Sequence length distribution

-ERX1461867, SRX9296596, SRX9296593, SRX9296590

The length of the sequence may differ depending on the sequencing platform. It will display the distribution of fragment sizes.

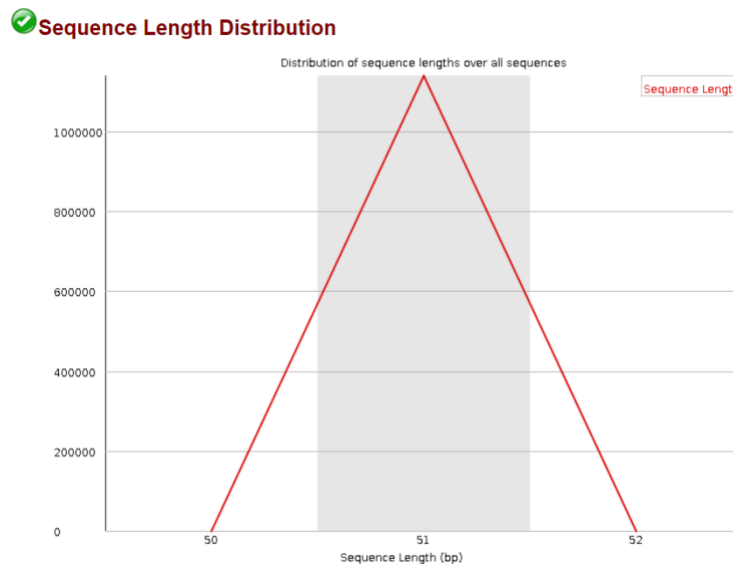


Fig:5.5 : Sequence length distribution of the run ERR1378585

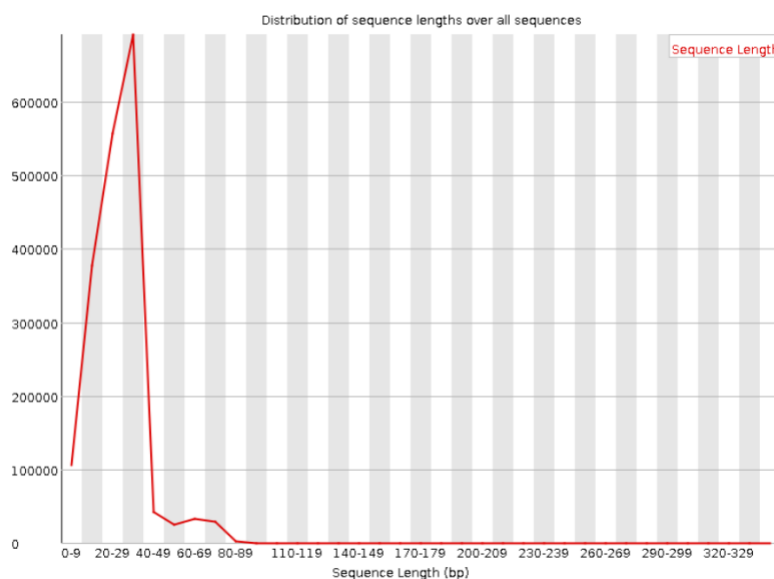


Fig:5.6 : Sequence length distribution of the run SRR12828917

Adapters and PCR fragments are removed by trimming after the raw sequence quality is determined, and the file is compressed and saved with .gz file extensions.

5.2 ALIGNMENT RESULTS

5.2.1 Bowtie2 alignment results

The Bowtie2 was used to align (mapping) the ERX1461867 genome raw sequence with the Human:hg38 genome. It generates a SAM-formatted output. @ is the SAM header line, with @SQ defining the alignment sorting order, SN defining the reference sequence dictionary, and LN defining the reference sequence length. There are 12 tabs that separate the fields (QNAME, FLAG, RNAME, POS, MAPQ, CIGAR, MRNM, MPOS, ISIZE, SEQ, QUAL, OPT). In Fig:5.7, you can see the chromosome number, position, and aligned reads.

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	MRNM	MPOS	ISIZE	SEQ	QUAL	OPT
@HD VN:1.0 SQ:coordinate											
@SQ SN:chr1 LN:248956422											
ERR1378585.560929	0	chr1	13373	1	51M	*		0	AAGTCACTCCTGCTTTTCCTTCCCTAGAGCCTCCACCCCGAGATC	BBBFFFFFFFFF	
ERR1378585.517638	0	chr1	19083	1	51M	*		0	GCACATCACCCACTCATGCTTTCCATGTTCTTGGCCGACGCAAGGCCG	BBBFFBFBFFFI	
ERR1378585.111227	0	chr1	19795	1	51M	*		0	TCCATCTTTTCTGGCTGGGAGAGGCTTCATCTGCTGTAAGGGTCTCTCC	BBBFFFFFFFFF	
ERR1378585.533522	0	chr1	22208	1	51M	*		0	TGCTGGACCTGACCTGTGTCATGTCACCTTGCTACCACGAGAGCATGGCC	BBBFFFFFFFFF	
ERR1378585.309717	0	chr1	25765	1	51M	*		0	CCTGGCACTGTGCCCTAGACCTGCTCCCTAGGCACTACAGTGGGGCCCTT	BBBFFFFFFFFF	
ERR1378585.789640	0	chr1	25765	1	51M	*		0	CCTGGCACTGTGCCCTAGACCTGCTCCCTAGGCACTACAGTGGGGCCCTT	BBBFFFFFFFFF	
ERR1378585.96750	16	chr1	25865	1	51M	*		0	AGGAGATTTCCTTGGGTTTAAATGCTGTGACCTTGAGTAAAGTTGCCGTC	BIIFIFIFFFFI	
ERR1378585.295521	0	chr1	25974	1	51M	*		0	CGTGGTATTTCCTAATAAATTGTATATTTTACATAACGCACACACCA	BBBFFFFFFFFF	
ERR1378585.773507	0	chr1	25974	1	51M	*		0	CGTGGTATTTCCTAATAAATTGTATATTTTACATAACGCACACACCA	BBBFFFFFFFFF	
ERR1378585.501429	16	chr1	31085	1	51M	*		0	GGGATAAATGATGATTTTAAAGTCGCCAGTGTGGTATTTTGTTTTGA	IIIIIIIIIIII	
ERR1378585.340148	16	chr1	32831	1	51M	*		0	CTCTGAGAGGCAAGAGAGCATAGTGGTACTAGGAAGGGAAGGACTTGGG	IIIIIIIIIIII	
ERR1378585.165558	16	chr1	34478	1	51M	*		0	CAGAGATAAACTTAAAGGCCAAGTAGGAGTTACCTGGTGAAGCAGGGCAG	FIIIIIIFBIIII	
ERR1378585.646055	16	chr1	34478	1	51M	*		0	CAGAGATAAACTTAAAGGCCAAGTAGGAGTTACCTGGTGAAGCAGGGCAG	7BFFB8B<FFB	
ERR1378585.9401	0	chr1	39398	1	51M	*		0	GATCTCAAGAGGGGTGAATGTTTGTGATGCACAGGACAGGAGTCTTG	BBBFFFFFFFFF	
ERR1378585.177018	16	chr1	51987	1	51M	*		0	GAACATGGGTGAGTAAGACTGAAGCAGCGTAATTACTGCACGATGCGCAT	IIIIIIIIIIII	
ERR1378585.499192	0	chr1	55739	1	51M	*		0	CTAACAGCCCTCAAAATTTTTCATTCACTGATGCTCTCTCATTTGCATTT	BBBFFFFFFFFF	
ERR1378585.268165	16	chr1	55863	1	51M	*		0	GCAACATGACTAATGATTTGAACGATGCTTACGCTTCTCTTACGTTCAA	IIIFIIIIIIII	
ERR1378585.426124	16	chr1	55863	1	51M	*		0	GCAACATGACTAATGATTTGAACGATGCTTACGCTTCTCTTACGTTCAA	IIIIIIIIIIII	
ERR1378585.48204	0	chr1	57570	1	51M	*		0	CTTTGGGGTATACATAAATGAAGATTAAGCTATCTGAATTCTCTCTCTC	BBBFFFFFFFFF	
ERR1378585.35932	0	chr1	62680	30	51M	*		0	AGTTTCTTAAGATGCTAATAAGTTTCATGATTTGAGCAATTGAGAACATT	BBBFFFFFFFFF	
ERR1378585.185627	0	chr1	71443	1	51M	*		0	GTGCAGGGGGGAAAGGCTTACTAATGTGATGTGGGATAATTGGATAGCAA	BBBFFFFFFFFF	
ERR1378585.700870	16	chr1	71550	1	51M	*		0	AGTGAATTCAAGACATACTCTTTTCAAAAAAATGCAATATATAAA	IIIIIIIIIIII	
ERR1378585.55180	16	chr1	73438	1	51M	*		0	TGTCTTTGCTATTGTGAATAGCACTGGGACAACCATACAAGTTTCATGTG	IIIIIIIIIIII	
ERR1378585.482935	16	chr1	73438	1	51M	*		0	TGTCTTTGCTATTGTGAATAGCACTGGGACAACCATACAAGTTTCATGTG	IIIIIIIIIIII	
ERR1378585.281393	0	chr1	78217	1	51M	*		0	AAGTATTGATTATCAATATAGACAATAAACCTTGAATGTACATGG	BBBFFFFFFFFF	

Fig:5.7 : Alignment of raw sequence (genome) with the reference genome hg38

```
882649 reads; of these:
  882649 (100.00%) were unpaired; of these:
    38278 (4.34%) aligned 0 times
    620015 (70.24%) aligned exactly 1 time
    224356 (25.42%) aligned >1 times
95.66% overall alignment rate
```

Fig:5.8 : Bowtie2 alignment statistics of ERR1378585

5.2.2 HISAT2

The RNA-seq reads are aligned (mapped) to the human: hg38 sequence as a reference. SAM format will be the output. When compared to TopHat2, it aligns very quickly and with good quality.

SRR12828917.538669	16	chr1	155089	0	57M95	*	0	0	TTCCCCGATATCACACTGGCTCTTCTCTACCTGTCTACCTGCTTAACCTAATAGGCCGTTACCA
SRR12828917.531844	272	chr1	187421	1	46M	*	0	0	CTGGAATGGTGCCAGGGGCAGAGGGGCAATGCCGGGGCCCCAGGTC
SRR12828917.1281698	16	chr1	258981	1	22M	*	0	0	TTCCGGCTTTAGGAGGAGCAGGA
SRR12828917.538669	272	chr1	513681	0	57M95	*	0	0	TTCCCCGATATCACACTGGCTCTTCTCTACCTGTCTACCTGCTTAACCTAATAGGCCGTTACCA
SRR12828917.1485308	256	chr1	629116	1	19M	*	0	0	GAATACGCCGAGGGCCCT
SRR12828917.717083	256	chr1	629222	1	50M	*	0	0	CACCTCTCCCTGAACTCTACACAACATATTTGTCCACCAAGACCCCTACTT
SRR12828917.715717	0	chr1	629305	1	22M	*	0	0	TACCCCGATTCCGCTACGACC
SRR12828917.906031	256	chr1	629335	1	26M	*	0	0	CACCTCCCTATGAAAAAATCTTCTTACC
SRR12828917.1564366	256	chr1	629347	0	4535M	*	0	0	TATGAAAACTTCTTACCCTACCCCTAGCATTACTTAT
SRR12828917.1666626	0	chr1	629382	1	25M	*	0	0	ATGATATGCTCCATCCCATTAACA
SRR12828917.63950	256	chr1	629433	1	39M	*	0	0	AGAAATATGCTGATAAAAGAGTTACTTTGATAGAGTAA
SRR12828917.79274	0	chr1	629433	1	42M	*	0	0	AGAAATATGCTGATAAAAGAGTTACTTTGATAGAGTAAATA
SRR12828917.171389	256	chr1	629433	0	40M	*	0	0	AGAAATATGCTGATAAAAGAGTTACTTTGATAGAGTAA
SRR12828917.258153	256	chr1	629433	0	39M	*	0	0	AGAAATATGCTGATAAAAGAGTTACTTTGATAGAGTAA
SRR12828917.274050	256	chr1	629433	0	39M	*	0	0	AGAAATATGCTGATAAAAGAGTTACTTTGATAGAGTAA
SRR12828917.292802	256	chr1	629433	0	49M	*	0	0	AGAAATATGCTGATAAAAGAGTTACTTTGATAGAGTAAATAATAGGAG
SRR12828917.315383	256	chr1	629433	0	39M	*	0	0	AGAAATATGCTGATAAAAGAGTTACTTTGATAGAGTAA
SRR12828917.382280	0	chr1	629433	0	46M	*	0	0	AGAAATATGCTGATAAAAGAGTTACTTTGATAGAGTAAATAATAG
SRR12828917.562650	256	chr1	629433	1	49M	*	0	0	AGAAATATGCTGATAAAAGAGTTACTTTGATAGAGTAAATAATAGGAG
SRR12828917.619461	256	chr1	629433	1	28M	*	0	0	AGAAATATGCTGATAAAAGAGTTACTT
SRR12828917.724751	0	chr1	629433	0	37M	*	0	0	AGAAATATGCTGATAAAAGAGTTACTTTGATAGAGT
SRR12828917.848203	256	chr1	629433	1	21M	*	0	0	AGAAATATGCTGATAAAAGA
SRR12828917.903069	256	chr1	629433	0	37M	*	0	0	AGAAATATGCTGATAAAAGAGTTACTTTGATAGAGT
SRR12828917.1120840	256	chr1	629433	1	20M	*	0	0	AGAAATATGCTGATAAAAG
SRR12828917.1188507	256	chr1	629433	0	44M	*	0	0	AGAAATATGCTGATAAAAGAGTTACTTTGATAGAGTAAATAAT
SRR12828917.1227285	256	chr1	629433	0	44M	*	0	0	AGAAATATGCTGATAAAAGAGTTACTTTGATAGAGTAAATAAT

Fig:5.9 : HISAT2 alignment report of raw sequence (transcriptome) with reference genome: hg38

```
HISAT2 summary stats:
Total reads: 1700661
  Aligned 0 time: 829209 (48.76%)
  Aligned 1 time: 275713 (16.21%)
  Aligned >1 times: 595739 (35.03%)
Overall alignment rate: 51.24%
```

Fig:5.10 : HISAT2 alignment statistics of run SRR12828917

5.3 SAMTOOLS RESULTS

This converts the SAM to BAM file by sorting the file.

ERR1378585.560929	0	chr1	13373	1	51M	*	0	0	AAGTTCACCTCCTGCCTTTTCCTTTCCCTAGAGCCTCCACCCCGAGATC
ERR1378585.517638	0	chr1	19083	1	51M	*	0	0	GCACATCACCCCACTCATGCTTTTCCATGTTCTTTGGCCGCGACGAAGGCCG
ERR1378585.111227	0	chr1	19795	1	51M	*	0	0	TCCATCTTTTCTGGCTGGGGAGAGGCCCTTCATCTGCTGTAAGGGTCTCC
ERR1378585.533522	0	chr1	22208	1	51M	*	0	0	TGCTGGACCTGACCCCTGTGTCATGTCACTTGTCTACCCAGAGAGCATGGCC
ERR1378585.309717	0	chr1	25765	1	51M	*	0	0	CCTGGCACTGTGCCCTAGACCTGTCTCCCTAGGCACACAGTGGGGCCCTT
ERR1378585.789640	0	chr1	25765	1	51M	*	0	0	CCTGGCACTGTGCCCTAGACCTGTCTCCCTAGGCACACAGTGGGGCCCTT
ERR1378585.96750	16	chr1	25865	1	51M	*	0	0	AGGAGATTTCACCTGGGTTTAAATGCTGTGACCTTGAGTAAGTGGCGTC
ERR1378585.295521	0	chr1	25974	1	51M	*	0	0	CGTGGTTATTCCAAATAAATTTGTATATTTTACATAACGCCACCAACCA
ERR1378585.773507	0	chr1	25974	1	51M	*	0	0	CGTGGTTATTCCAAATAAATTTGTATATTTTACATAACGCCACCAACCA
ERR1378585.501429	16	chr1	31085	1	51M	*	0	0	GGGATAAATGTATGATTTTAAAGTCGCCAGTGTGTGATTTTGTTTTGA
ERR1378585.340148	16	chr1	32831	1	51M	*	0	0	CTCTGAGAGGCAGGAGACATAGTGGTTACTAGGAAGGGAAGGACTTTGGG
ERR1378585.165558	16	chr1	34478	1	51M	*	0	0	CAGAGATAAACTTAAAGGCCAAGTAGGAGTTACCTCGTGAAGCAGGGCAC
ERR1378585.646055	16	chr1	34478	1	51M	*	0	0	CAGAGATAAACTTAAAGGCCAAGTAGGAGTTACCTCGTGAAGCAGGGCAC
ERR1378585.9401	0	chr1	39398	1	51M	*	0	0	GATCTCAAGAGGGGTGAATGTGTTTTCATGCAAGAGGACAGGAGTCTTG
ERR1378585.177018	16	chr1	51987	1	51M	*	0	0	GAACATGGGTGAGTAAGACTGAAGCAGCCGTAATTAAGTGCACGATGCGCAT
ERR1378585.499192	0	chr1	55739	1	51M	*	0	0	CTAACAGCCCTCAAAATTTTCATTTCAGTGATGCTCTCTTCATTGCTTTT
ERR1378585.268165	16	chr1	55863	1	51M	*	0	0	GCAACATGACTAATGTATTGAACGATGCTTCAGCTTCTCTTCACGTTCAA
ERR1378585.426124	16	chr1	55863	1	51M	*	0	0	GCAACATGACTAATGTATTGAACGATGCTTCAGCTTCTCTTCACGTTCAA
ERR1378585.48204	0	chr1	57570	1	51M	*	0	0	CTTTGGGGTATTACATAAATGAAGATTAAGCTACTGAATTTCTCTCTCTC
ERR1378585.35932	0	chr1	62680	30	51M	*	0	0	AGTTTCTTAAGATGCTAAATAGTTTCATGATTTGAGCAATTTGAGAATCTT
ERR1378585.185627	0	chr1	71443	1	51M	*	0	0	GTGCAGGGGGGAAAGGCCCTACTAATGTATGATGGTGGGATAATTTGATAGCAA
ERR1378585.700870	16	chr1	71550	1	51M	*	0	0	AGTGAAATCAAGACATAAATCTTTTTCACAAAAAATATGACAAATATAAA
ERR1378585.55180	16	chr1	73438	1	51M	*	0	0	TGCTTTGCTATTGTGAATAGCAGTGGGACAAACCCATCAAGTTCATGTGT
ERR1378585.482935	16	chr1	73438	1	51M	*	0	0	TGCTTTGCTATTGTGAATAGCAGTGGGACAAACCCATCAAGTTCATGTGT
ERR1378585.281393	0	chr1	78217	1	51M	*	0	0	AAGTATTGATATATCAATATAGACAAATAAATCTTGAATGTACATGG
ERR1378585.243278	0	chr1	82364	1	51M	*	0	0	GGAAAAAGCAGGAGAGAAAGTTCTCTTTGGGAAATCGAATCACTCAAAAGTC
ERR1378585.240208	0	chr1	87062	32	51M	*	0	0	TAGAAGAGATTTTAAAAACATGACAAATGAAAAAATGGGCAAGACTAAA
ERR1378585.434324	0	chr1	87062	32	51M	*	0	0	TAGAAGAGATTTTAAAAACATGACAAATGAAAAAATGGGCAAGACTAAA
ERR1378585.642836	0	chr1	87062	32	51M	*	0	0	TAGAAGAGATTTTAAAAACATGACAAATGAAAAAATGGGCAAGACTAAA
ERR1378585.860721	0	chr1	87062	32	51M	*	0	0	TAGAAGAGATTTTAAAAACATGACAAATGAAAAAATGGGCAAGACTAAA
ERR1378585.750796	0	chr1	99957	1	51M	*	0	0	TGTGAGAAAGGATGAATCTTCTGAAGGTTATGTCATCACTCACTAAAGC
ERR1378585.123813	0	chr1	101715	1	51M	*	0	0	ACAATAAAATAAATGATTAAAGTATGAACACAAAAAAGCTAGTAAAGGT
ERR1378585.202859	16	chr1	103585	1	51M	*	0	0	TTGGTGCCATCAAGTGAGTCTTGCTCGCTCTTTTCTTTTGTGAT

Fig:5.11 : Samtools filtered alignments of Genome (BAM)

5.4 VARIANT CALLING

Variant calling is used to identify variants or SNPs. The file format for the output is a VCF file. Read depth, allele frequency, Phred-scaled strand bias, and counts for ref-forward, ref-reverse, alt-forward, and alt-reverse bases are all included. It contains the nucleotide that has been changed in the sequence.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
chr3	93470433	.	A	T	142	PASS	DP=80;AF=0.550000;SB=7;DP4=23,13,21,23
chr3	93470437	.	T	G	303	PASS	DP=73;AF=0.671233;SB=13;DP4=18,6,24,25
chr3	93470446	.	G	A	872	PASS	DP=79;AF=0.949367;SB=2;DP4=4,0,55,20
chr3	93470523	.	G	A	1883	PASS	DP=130;AF=0.984615;SB=0;DP4=1,1,72,56
chr3	93470569	.	T	A	2846	PASS	DP=156;AF=0.955128;SB=1;DP4=4,3,72,77
chr3	93470628	.	A	C	348	PASS	DP=75;AF=0.720000;SB=5;DP4=10,7,23,31
chr3	93470632	.	T	C	220	PASS	DP=69;AF=0.811594;SB=6;DP4=9,3,28,28
chr3	93470646	.	G	A	264	PASS	DP=76;AF=0.921053;SB=1;DP4=4,2,36,34
chr3	93470689	.	C	G	2388	PASS	DP=177;AF=0.858757;SB=0;DP4=18,7,110,42
chr3	93470771	.	C	G	84	PASS	DP=49;AF=0.408163;SB=3;DP4=7,21,3,17
chr3	93470789	.	G	T	78	PASS	DP=19;AF=0.842105;SB=29;DP4=3,0,0,16
chr5	49601530	.	A	T	139	PASS	DP=16;AF=1.000000;SB=0;DP4=0,0,10,6
chr5	49601573	.	A	G	96	PASS	DP=14;AF=0.785714;SB=3;DP4=2,0,6,5
chr10	133689420	.	G	T	117	PASS	DP=10;AF=0.800000;SB=0;DP4=0,2,0,8
chrUn_K1270438v1	108796	.	A	T	247	PASS	DP=16;AF=1.000000;SB=0;DP4=0,0,5,11
chrUn_K1270438v1	108800	.	A	G	186	PASS	DP=18;AF=0.888889;SB=0;DP4=0,2,5,11
chrUn_K1270438v1	109425	.	G	A	205	PASS	DP=10;AF=1.000000;SB=0;DP4=0,0,5,5
chrUn_K1270438v1	112093	.	A	T	100	PASS	DP=13;AF=1.000000;SB=0;DP4=0,0,11,2
chrUn_K1270467v1	2928	.	G	T	238	PASS	DP=10;AF=1.000000;SB=0;DP4=0,0,6,4
chrUn_K1270467v1	2945	.	C	G	238	PASS	DP=10;AF=1.000000;SB=0;DP4=0,0,6,4

Fig:5.12 : Call Variants of genome (vcf)

#CHROM	POS	ID	REF	ALT
chr1	4198187	.	G	T
chr1	4331760	rs1691447	T	C
chr1	4822596	rs1883345	G	A
chr1	5218990	rs4847555	G	A
chr1	6525744	rs13513	C	T
chr1	8989320	.	G	A
chr1	9300300	rs57478549	C	A
chr1	15974660	.	C	T
chr1	16064493	rs1414500	G	A
chr1	16206763	rs221055	C	T
chr1	20297688	.	G	A
chr1	24177849	rs55976784	G	C
chr1	24890041	rs6663476	T	C
chr1	25843053	.	C	T
chr1	28875694	.	A	C
chr1	32069644	.	G	A
chr1	32254007	.	A	T
chr1	38475515	.	G	T
chr1	39635841	rs1148941	C	T
chr1	41779381	.	C	T
chr1	42084906	rs11802965	T	C
chr1	44013071	.	C	T
chr1	44207074	.	C	T
chr1	46350972	rs57267378	A	G
chr1	49139268	rs617996	A	C
chr1	49988063	rs6685494	A	G
chr1	50966703	.	G	A
chr1	52730128	rs928451	A	G
chr1	53291659	rs1288501	A	C
chr1	53345668	.	C	T
chr1	53755563	rs10888808	G	T
chr1	53755564	rs10888809	T	G

Fig:5.13 : SnpSift Annotation of genome(rs id's of SNIPS)

The SNPs with their reference IDs are shown in VCF file format in Fig:5.12. The rs Id will be generated by combining the annotation data. The file contains 5,400 lines, each with the following information: chromosome where the variation occurred, position where the variation occurred, rs ID (Ex: rs1691447), altered nucleotide in the sequence, minimum SNP quality, allele frequency, Phred-scaled strand bias at this position. Bases for ref-forward, ref-reverse, alt-forward, and alt-reverse are all counted.

5.5 ASSEMBLED READS

5.5.1. Cufflink

tracking_id	class_code	nearest_ref_id	gene_id	gene_short_name	tss_id	locus	length	coverage	FPKM	FPKM_conf_lo	FPKM_conf_hi	FPKM_status
CUFF.1.1	-	-	CUFF.1	-	-	chr1:9151691-9151760	69	12726.1	648147	588414	707880	OK
CUFF.2.1	-	-	CUFF.2	-	-	chr1:28508702-28508762	60	1236.69	39389.8	20809	57970.5	OK
CUFF.3.1	-	-	CUFF.3	-	-	chr1:28648599-28648733	134	77.3041	2496.95	1285.75	3708.15	OK
CUFF.4.1	-	-	CUFF.4	-	-	chr1:28578751-28578821	70	6509.43	132628	106282	158974	OK
CUFF.5.1	-	-	CUFF.5	-	-	chr1:30968164-30968235	71	622.043	14879.5	6256.25	23502.8	OK
CUFF.6.1	-	-	CUFF.6	-	-	chr1:40754370-40754434	64	28740.6	1.43251e+06	1.33147e+06	1.53355e+06	OK
CUFF.7.1	-	-	CUFF.7	-	-	chr1:44776491-44776592	101	371.402	5542.95	2581.88	8504.01	OK
CUFF.8.1	-	-	CUFF.8	-	-	chr1:44777842-44777910	68	961.856	20062.5	9338.63	30786.3	OK
CUFF.9.1	-	-	CUFF.9	-	-	chr1:44775863-44775938	75	2814.13	53687.5	38787.2	68587.7	OK
CUFF.10.1	-	-	CUFF.10	-	-	chr1:44778389-44778456	67	20346	411187	360789	461584	OK
CUFF.11.1	-	-	CUFF.11	-	-	chr1:71067640-71067704	64	6997.57	372628	320762	424495	OK
CUFF.12.1	-	-	CUFF.12	-	-	chr1:73806949-73806987	38	16914	630151	458592	801709	OK
CUFF.13.1	-	-	CUFF.13	-	-	chr1:75787887-75787971	84	1211.86	24767.2	16294.8	33239.7	OK
CUFF.14.1	-	-	CUFF.14	-	-	chr1:75787070-75787149	79	4898.48	83347.2	66334	100360	OK
CUFF.15.1	-	-	CUFF.15	-	-	chr1:75789475-75789547	72	9036.67	173548	144855	202241	OK
CUFF.16.1	-	-	CUFF.16	-	-	chr1:92837288-92837378	90	380.78	10682	5773.86	15590.2	OK
CUFF.17.1	-	-	CUFF.17	-	-	chr1:95155792-95155819	27	69010.4	3.41149e+06	2.49172e+06	4.33126e+06	OK

Fig:5.14 : Cufflinks: the transcripts are assembled

5.5.2 Cuffmerge

The transcripts that are assembled before are merged into mega transcript. This step is required for DEG analysis.

chr1	Cufflinks	exon	9151692	9151760	.	.	.	gene_id "XLOC_000001"; transcript_id "TCONS_00000001"; exon_number "1"; old "CUFF.3.1"; tss_id "TSS1";
chr1	Cufflinks	exon	28508703	28508762	.	.	.	gene_id "XLOC_000002"; transcript_id "TCONS_00000002"; exon_number "1"; old "CUFF.1.1"; tss_id "TSS2";
chr1	Cufflinks	exon	28578752	28578821	.	.	.	gene_id "XLOC_000003"; transcript_id "TCONS_00000003"; exon_number "1"; old "CUFF.6.1"; tss_id "TSS3";
chr1	Cufflinks	exon	28648600	28648733	.	.	.	gene_id "XLOC_000004"; transcript_id "TCONS_00000004"; exon_number "1"; old "CUFF.2.1"; tss_id "TSS4";
chr1	Cufflinks	exon	30968165	30968235	.	.	.	gene_id "XLOC_000005"; transcript_id "TCONS_00000005"; exon_number "1"; old "CUFF.5.1"; tss_id "TSS5";
chr1	Cufflinks	exon	40754371	40754434	.	.	.	gene_id "XLOC_000006"; transcript_id "TCONS_00000006"; exon_number "1"; old "CUFF.4.1"; tss_id "TSS6";
chr1	Cufflinks	exon	44775864	44775938	.	.	.	gene_id "XLOC_000007"; transcript_id "TCONS_00000007"; exon_number "1"; old "CUFF.7.1"; tss_id "TSS7";
chr1	Cufflinks	exon	44776492	44776592	.	.	.	gene_id "XLOC_000008"; transcript_id "TCONS_00000008"; exon_number "1"; old "CUFF.8.1"; tss_id "TSS8";
chr1	Cufflinks	exon	44777843	44777910	.	.	.	gene_id "XLOC_000009"; transcript_id "TCONS_00000009"; exon_number "1"; old "CUFF.9.1"; tss_id "TSS9";
chr1	Cufflinks	exon	44778390	44778456	.	.	.	gene_id "XLOC_000010"; transcript_id "TCONS_00000010"; exon_number "1"; old "CUFF.10.1"; tss_id "TSS10";
chr1	Cufflinks	exon	71067641	71067704	.	.	.	gene_id "XLOC_000011"; transcript_id "TCONS_00000011"; exon_number "1"; old "CUFF.11.1"; tss_id "TSS11";
chr1	Cufflinks	exon	73806950	73806987	.	.	.	gene_id "XLOC_000012"; transcript_id "TCONS_00000012"; exon_number "1"; old "CUFF.12.1"; tss_id "TSS12";
chr1	Cufflinks	exon	75787071	75787149	.	.	.	gene_id "XLOC_000013"; transcript_id "TCONS_00000013"; exon_number "1"; old "CUFF.13.1"; tss_id "TSS13";
chr1	Cufflinks	exon	75787888	75787971	.	.	.	gene_id "XLOC_000014"; transcript_id "TCONS_00000014"; exon_number "1"; old "CUFF.14.1"; tss_id "TSS14";
chr1	Cufflinks	exon	75789476	75789547	.	.	.	gene_id "XLOC_000015"; transcript_id "TCONS_00000015"; exon_number "1"; old "CUFF.15.1"; tss_id "TSS15";
chr1	Cufflinks	exon	92837289	92837378	.	.	.	gene_id "XLOC_000016"; transcript_id "TCONS_00000016"; exon_number "1"; old "CUFF.16.1"; tss_id "TSS16";
chr1	Cufflinks	exon	93847573	93847662	.	.	.	gene_id "XLOC_000017"; transcript_id "TCONS_00000017"; exon_number "1"; old "CUFF.17.1"; tss_id "TSS17";
chr1	Cufflinks	exon	95155793	95155819	.	.	.	gene_id "XLOC_000018"; transcript_id "TCONS_00000018"; exon_number "1"; old "CUFF.18.1"; tss_id "TSS18";
chr1	Cufflinks	exon	98046090	98046114	.	.	.	gene_id "XLOC_000019"; transcript_id "TCONS_00000019"; exon_number "1"; old "CUFF.19.1"; tss_id "TSS19";
chr1	Cufflinks	exon	109100536	109100612	.	.	.	gene_id "XLOC_000020"; transcript_id "TCONS_00000020"; exon_number "1"; old "CUFF.20.1"; tss_id "TSS20";
chr1	Cufflinks	exon	109598940	109598963	.	.	.	gene_id "XLOC_000021"; transcript_id "TCONS_00000021"; exon_number "1"; old "CUFF.23.1"; tss_id "TSS21";
chr1	Cufflinks	exon	121016855	121016918	.	.	.	gene_id "XLOC_000022"; transcript_id "TCONS_00000022"; exon_number "1"; old "CUFF.25.1"; tss_id "TSS22";
chr1	Cufflinks	exon	145287766	145287841	.	.	.	gene_id "XLOC_000023"; transcript_id "TCONS_00000023"; exon_number "1"; old "CUFF.21.1"; tss_id "TSS23";
chr1	Cufflinks	exon	155195196	155195258	.	.	.	gene_id "XLOC_000024"; transcript_id "TCONS_00000024"; exon_number "1"; old "CUFF.22.1"; tss_id "TSS24";
chr1	Cufflinks	exon	159141588	159141685	.	.	.	gene_id "XLOC_000025"; transcript_id "TCONS_00000025"; exon_number "1"; old "CUFF.24.1"; tss_id "TSS25";

Fig:5.15 Merged transcriptome

5.6 DIFFERENTIALLY REGULATING GENES AND BASE STATISTICS

This table contains information about DEGs such as their ID, adj.p-value, p. value, t, B, logFC, gene symbol, gene title, gene id, and chromosome location. The p-values of all 46 genes are ranked, with the genes with the lowest p values being the most significant. CHRN2, TPI1, NaN, SYT5, ORAI2, PSMD8, EHD3, INPP5A, TTPAL, CLIP3, CPT1C, NACC1, LINC00643, NUDT14, SYNGR3, RAB39B, NRN1, TTYH3, CCKBR, FAM102B, TMEM183B/TMEM183A, VLDLR, RAB15, HPCA, NUS1... are among the novel genes listed in 5.1. These genes are expressed differently, and genes that are present in this table but not in the PhenGenI AD are false positives.

Table:5.1 Differentially regulating genes

	ID	ADJ.P. VALUE	P.VAL UE	t	B	logF C	GENE. SYMBOL	GENE. TITLE	GENE. ID	
0	790 581 7	4.2900 00e-07	1.2900 00e-11	- 7.86 345 7	15.98 1649	- 0.40 935 0	CHRNA2	cholinergic receptor nicotinic beta 2 subunit	1141	
1	818 037 1	5.3200 00e-07	5.3500 00e-11	- 7.55 010 2	14.64 5976	- 0.39 336 8	TPI1	triosephosp hate isomerase 1	7167	
2	812 113 0	5.3200 00e-07	6.0800 00e-11	7.52 164 2	14.52 4991	0.47 819 9	NaN	NaN	NaN	
3	803 937 8	5.3200 00e-07	6.3900 00e-11	- 7.51 072 8	14.47 8610	- 0.46 258 9	SYT5	synaptotag min 5	6861	
4	789 380 8	1.0100 00e-06	1.6900 00e-10	- 7.29 538 0	13.56 5555	- 0.32 866 6	NaN	NaN	NaN	
5	813 517 2	1.0100 00e-06	1.8300 00e-10	- 7.27 816 1	13.49 2732	- 0.50 103 4	ORAI2	ORAI calcium release- activated calcium modulat...	80228	
6	802 838 0	1.7900 00e-06	3.8900 00e-10	- 7.10 988 5	12.78 2699	- 0.31 749 7	PSMD8	proteasome 26S subunit, non-	5714	

								ATPase 8		
7	789 459 6	1.7900 00e-06	4.2900 00e-10	- 7.08 808 6	12.69 0952	- 0.34 677 4	NaN	NaN	NaN	
8	804 122 5	2.6500 00e-06	7.1500 00e-10	- 6.97 385 6	12.21 1114	- 0.49 131 0	EHD3	EH domain containing 3	30845	
9	806 279 6	2.7300 00e-06	8.9300 00e-10	- 6.92 415 0	12.00 2842	- 0.29 780 8	GDAP1L1	ganglioside induced differenti on associated.. .	78997	
10	793 147 9	2.7300 00e-06	9.0100 00e-10	- 6.92 217 9	11.99 4587	- 0.32 031 4	INPP5A	inositol polyphosph ate-5- phosphatas e A	3632	
11	806 284 4	3.2600 00e-06	1.1700 00e-09	- 6.86 274 1	11.74 5992	- 0.43 254 9	TTPAL	alpha tocopherol transfer protein like	79183	
12	798 256 4	3.3500 00e-06	1.3900 00e-09	- 6.82 501 0	11.58 8437	- 0.44 148 9	SPRED1	sprouty related EVH1 domain containing 1	161742	
13	803 625 2	3.3500 00e-06	1.4100 00e-09	- 6.82 173	11.57 4768	- 0.30 365	CLIP3	CAP-Gly domain containing	25999	

				4		9		linker protein 3		
14	803 044 8	4.8200 00e-06	2.1700 00e-09	- 6.72 417 4	11.16 8414	- 0.32 493 0	CPT1C	carnitine palmitoyltra nsferase 1C	126129	
15	802 615 5	5.2700 00e-06	2.5300 00e-09	- 6.68 936 8	11.02 3797	- 0.28 698 6	NACC1	nucleus accumbens associated 1	112939	
16	797 489 5	6.2200 00e-06	3.2900 00e-09	- 6.62 986 1	10.77 6995	- 0.62 189 3	LINC00643	long intergenic non-protein coding RNA 643	646113	
17	798 156 6	6.2200 00e-06	3.4200 00e-09	- 6.62 127 3	10.74 1426	- 0.24 460 2	NUDT14	nudix hydrolase 14	256281	
18	799 244 7	6.2200 00e-06	3.5500 00e-09	- 6.61 289 7	10.70 6746	- 0.32 339 2	SYNGR3	synaptogyri n 3	9143	
19	817 623 0	8.5400 00e-06	5.1300 00e-09	- 6.52 932 0	10.36 1360	- 0.33 950 8	RAB39B	RAB39B, member RAS oncogene family	116442	
20	812 373 9	8.7800 00e-06	5.7900 00e-09	- 6.50 155 5	10.24 6893	- 0.54 574 4	NRN1	neuritin 1	51299	

21	813 117 9	8.7800 00e-06	6.0100 00e-09	- 6.49 332 0	10.21 2968	- 0.40 086 2	TTYH3	tweety family member 3	80727	
22	793 809 0	8.7800 00e-06	6.0800 00e-09	- 6.49 047 4	10.20 1248	- 0.41 265 3	CCKBR	cholecystok inin B receptor	887	
23	789 418 5	8.7800 00e-06	6.3300 00e-09	- 6.48 129 6	10.16 3459	- 0.30 937 2	NaN	NaN	NaN	
24	790 350 7	1.0100 00e-05	7.6100 00e-09	- 6.43 923 5	9.990 474	- 0.54 638 5	FAM102B	family with sequence similarity 102 member B	284611	
25	809 145 2	1.0400 00e-05	8.1300 00e-09	- 6.42 410 5	9.928 333	- 0.31 295 4	TMEM183B///T MEM183A	transmembr ane protein 183B///trans membrane pro...	653659///92703	
26	815 410 0	1.1700 00e-05	9.4500 00e-09	- 6.38 977 3	9.787 482	- 0.33 200 3	VLDLR	very low density lipoprotein receptor	7436	
27	797 966 3	1.2600 00e-05	1.0800 00e-08	- 6.35 839 0	9.658 929	- 0.49 853 3	RAB15	RAB15, member RAS oncogene family	376267	
28	789 984 1	1.2600 00e-05	1.1300 00e-08	- 6.34 925 9	9.621 562	- 0.60 708 5	HPCA	hippocalcin	3208	

29	812 170 4	1.2600 00e-05	1.1300 00e-08	- 6.34 786 1	9.615 842	- 0.28 907 8	NUS1	NUS1 dehydrodoli chyl diphosphat e synthase subunit	116150	
...	...	...	...	...	...	...	...	...	...	
33 26 7	808 854 6	9.9900 00e-01	9.9800 00e-01	- 0.00 247 8	- 6.274 354	- 0.00 013 4	NaN	NaN	NaN	
33 26 8	789 262 9	9.9900 00e-01	9.9800 00e-01	0.00 246 4	- 6.274 354	0.00 023 5	NaN	NaN	NaN	
33 26 9	812 136 5	9.9900 00e-01	9.9800 00e-01	- 0.00 236 1	- 6.274 355	- 0.00 012 9	FOXO3	forkhead box O3	2309	
33 27 0	810 776 9	9.9900 00e-01	9.9800 00e-01	0.00 223 1	- 6.274 355	0.00 011 1	SLC12A2	solute carrier family 12 member 2	6558	
33 27 1	814 704 9	9.9900 00e-01	9.9800 00e-01	0.00 203 2	- 6.274 355	0.00 019 5	FABP5P3///FAB P5	fatty acid binding protein 5 pseudogen e 3///fa...	220832///2171	
33 27 2	793 701 6	9.9900 00e-01	9.9800 00e-01	- 0.00 193 0	- 6.274 355	- 0.00 009 0	CLRN3	clarin 3	119467	
33 27 3	793 249 2	9.9900 00e-01	9.9800 00e-01	- 0.00 189	- 6.274 356	- 0.00 011	C10orf113	chromosom e 10 open reading	387638	

				9		6		frame 113		
33 27 4	789 273 1	9.9900 00e-01	9.9900 00e-01	0.00 187 6	- 6.274 356	0.00 016 0	NaN	NaN	NaN	
33 27 5	793 611 5	9.9900 00e-01	9.9900 00e-01	- 0.00 183 1	- 6.274 356	- 0.00 009 1	SH3PXD2A	SH3 and PX domains 2A	9644	
33 27 6	795 248 4	9.9900 00e-01	9.9900 00e-01	0.00 177 6	- 6.274 356	0.00 011 5	TMEM218	transmembr ane protein 218	219854	
33 27 7	809 199 1	9.9900 00e-01	9.9900 00e-01	0.00 169 9	- 6.274 356	0.00 007 3	MECOM	MDS1 and EVI1 complex locus	2122	
33 27 8	817 360 5	9.9900 00e-01	9.9900 00e-01	0.00 168 1	- 6.274 356	0.00 009 2	NaN	NaN	NaN	
33 27 9	799 458 2	9.9900 00e-01	9.9900 00e-01	- 0.00 167 0	- 6.274 356	- 0.00 006 6	SLX1B- SULT1A4///SLX 1A- SULT1A3///SLX 1A///SULT1A...	SLX1B- SULT1A4 readthrough (NMD candidate)// /SL...	100526831///100526830///5 48593///445329///7900...	16p///16p///1 11.2///16p
33 28 0	814 906 7	9.9900 00e-01	9.9900 00e-01	- 0.00 160 8	- 6.274 356	- 0.00 008 4	NaN	NaN	NaN	
33 28 1	803 168 4	9.9900 00e-01	9.9900 00e-01	0.00 150 2	- 6.274 356	0.00 014 0	NaN	NaN	NaN	

33 28 2	797 013 9	9.9900 00e-01	9.9900 00e-01	0.00 146 6	- 6.274 356	0.00 005 9	TEX29	testis expressed 29	121793	
33 28 3	789 342 9	9.9900 00e-01	9.9900 00e-01	0.00 144 5	- 6.274 356	0.00 017 0	NaN	NaN	NaN	
33 28 4	800 490 5	9.9900 00e-01	9.9900 00e-01	0.00 142 0	- 6.274 356	0.00 008 1	USP43	ubiquitin specific peptidase 43	124739	
33 28 5	802 625 0	9.9900 00e-01	9.9900 00e-01	- 0.00 138 0	- 6.274 356	- 0.00 005 2	DCAF15	DDB1 and CUL4 associated factor 15	90379	
33 28 6	809 541 2	9.9900 00e-01	9.9900 00e-01	0.00 111 0	- 6.274 357	0.00 007 9	CSN1S1	casein alpha s1	1446	
33 28 7	802 734 8	9.9900 00e-01	9.9900 00e-01	0.00 108 4	- 6.274 357	0.00 013 1	ZNF730	zinc finger protein 730	100129543	
33 28 8	791 678 1	9.9900 00e-01	9.9900 00e-01	- 0.00 100 9	- 6.274 357	- 0.00 004 2	NaN	NaN	NaN	
33 28 9	804 190 0	1.0000 00e+00	9.9900 00e-01	0.00 079 0	- 6.274 357	0.00 003 6	NaN	NaN	NaN	
33 29 0	804 682 2	1.0000 00e+00	1.0000 00e+00	- 0.00 058 3	- 6.274 357	- 0.00 004 9	NaN	NaN	NaN	

33 29 1	799 866 6	1.0000 00e+00	1.0000 00e+00	- 0.00 057 2	- 6.274 357	- 0.00 004 7	SNORA64///RP S2	small nucleolar RNA, H/ACA box 64///riboso mal ...	26784///6187	16p1
33 29 2	801 062 9	1.0000 00e+00	1.0000 00e+00	- 0.00 056 0	- 6.274 357	- 0.00 002 0	CCDC137	coiled-coil domain containing 137	339230	
33 29 3	808 062 1	1.0000 00e+00	1.0000 00e+00	0.00 037 4	- 6.274 357	0.00 002 2	CCDC66	coiled-coil domain containing 66	285331	
33 29 4	789 531 3	1.0000 00e+00	1.0000 00e+00	- 0.00 036 1	- 6.274 357	- 0.00 003 1	NaN	NaN	NaN	
33 29 5	799 450 6	1.0000 00e+00	1.0000 00e+00	- 0.00 010 3	- 6.274 357	- 0.00 000 3	NFATC2IP	nuclear factor of activated T- cells 2 interact...	84901	
33 29 6	802 737 7	1.0000 00e+00	1.0000 00e+00	- 0.00 009 9	- 6.274 357	- 0.00 000 4	NaN	NaN		

5.7 SNPs THAT ARE EXPRESSED IN BOTH GENOTYPE AND PHENOTYPE

Table:5.2 List of SNPs (rs ID's) expressed in genotype and phenotype of AD

rs2970989	rs6665019	rs1347297
rs4676049	rs10800664	rs12691765
rs75286889	rs17725296	rs17366218
rs140009341	rs34487851	rs3820443
rs6678275	rs6738181	rs1357692
rs982100	rs12691765	rs340849
rs72842919	rs1340425	rs2421847
rs36060340	rs17125439	rs79036927
rs62179067	rs41458646	rs41458646
rs79036927	rs113884973	rs10800664
rs2121433	rs12134133	rs10925500
rs4499362	rs11889338	rs2244526
rs4670766	rs1357692	rs62177277
rs4267554	rs36060340	rs62179067
rs2969775	rs1340425	rs7518943
rs6733839	rs78022502	rs1857353
rs1125355	rs4400585	rs75847378
rs6733839	rs7539409	rs72840936
rs744373	rs11166407	rs17725296
rs7561528	rs6677080	rs115078773
rs4663105	rs113884973	rs12733952
rs7561528	rs6665019	rs76854344
rs12989701	rs982100	rs12044355
rs76854344	rs2970989	rs140009341
rs6733839	rs34487851	rs4038131
rs75286889	rs7588354	rs6714710
		rs178585

List of SNPs (rs ID's) expressed in genotype and phenotype of AD The PhenGenI and IGAP meta-analyses of the AD genome yielded the common SNPs that are expressed between genotype and phenotype. There are a total of 82 SNPs that are shared between the Phen-Gen and AD.

5.8 GO ANNOTATION

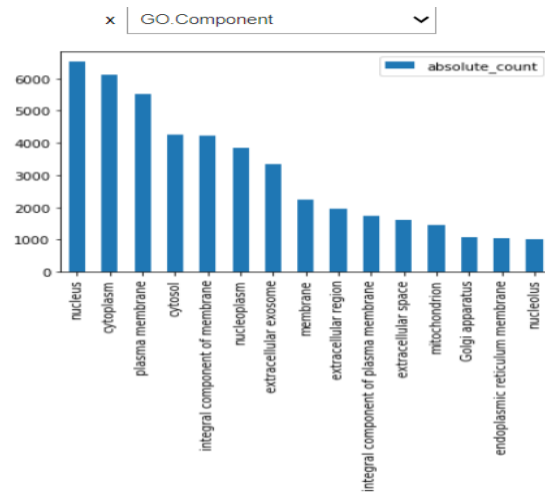


Fig:5.16 : GO Component

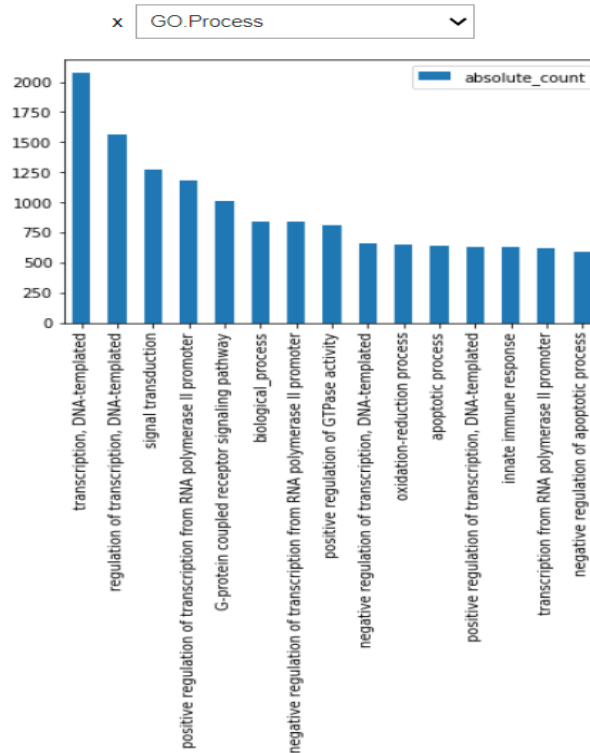


Fig: 5.17 : GO Process

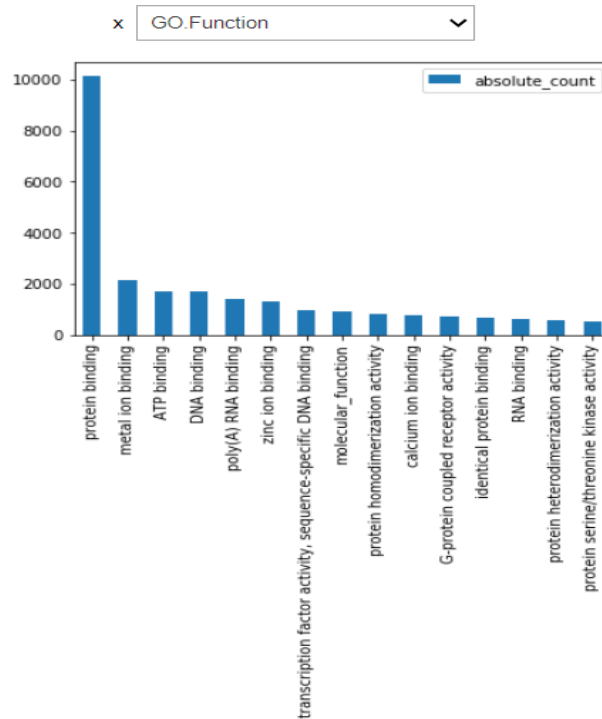


Fig:5.18 : GO Function

Annotation is visualised, compared, and plotted using these plots. The three different types of results for single samples are G.O function, which describes the functional actives, G.O process, which describes the processing, and G.O component, which describes the type of components or organs.

5.9 VOLCANO PLOT

This was made with Limma. A volcano plot depicts statistical significance ($-\log_{10}$ P value) versus magnitude of change ($\log_2$ fold change) and can be used to visualise differentially expressed genes data points as well as see individual gene annotation. The highlighted genes are significantly differentially expressed (red = upregulated, blue = downregulated) at a default adjusted p-value cut off of 0.05. Alzheimer's disease vs. non-disease Alzheimer's is compared here.

Volcano plot
GSE36980: Expression data from post mortem
Alzheimer's disease brains
AD vs NonAD, Padj<0.05

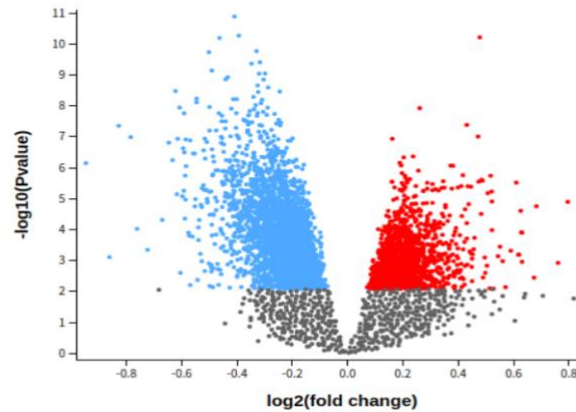


Fig: 5.19 Up-regulated & Down-regulated genes in AD Vs Non-AD

5.10 HEAT MAP

The heatmap is combined with the clustering method to group genes and samples together based on gene expression similarities. The gene is represented by the row, and the sample is represented by the column. This was discovered through a variety of visualisations of disease state, tissue, and gender. This shows the expression of the gens in all the samples. Row represents genes and column represents samples.

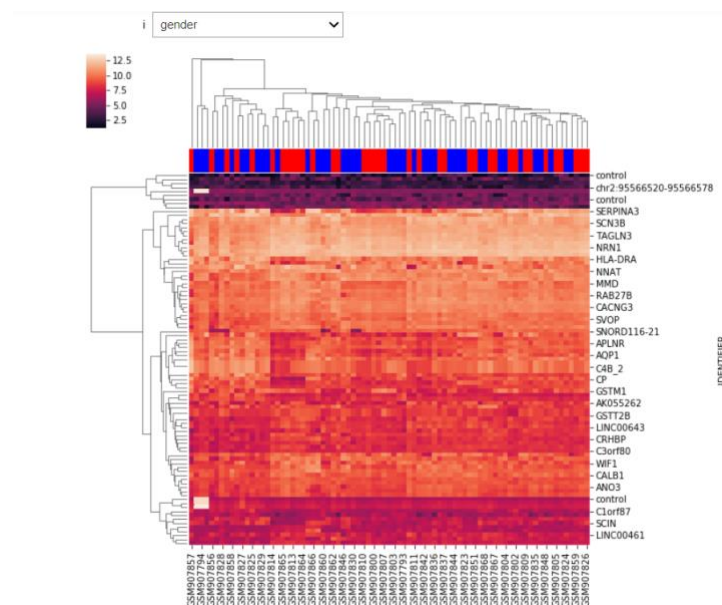


Fig:5.20 : Clustering of expressed genes based on gender

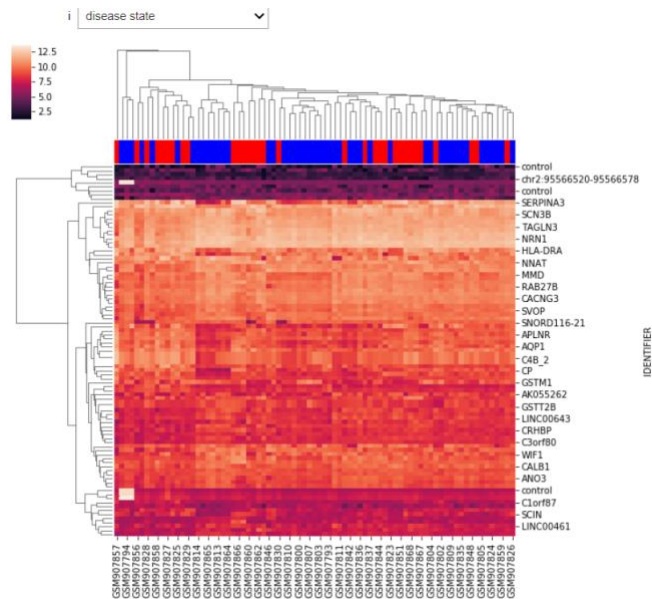


Fig:5.21 : Clustering of expressed genes based on disease state

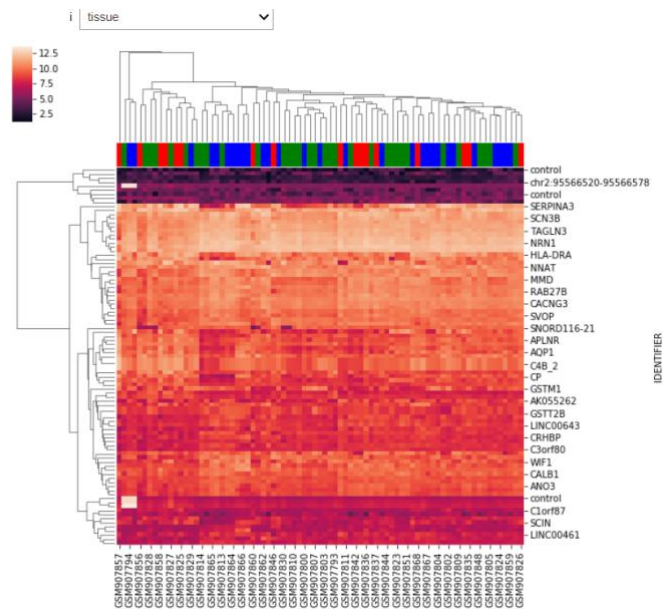


Fig:5.22 : Clustering of expressed genes based on Tissue

5.11 BOX PLOT

The gene expression was discovered using a box plot. There are two types of gene expression here: Alzheimer's disease and non-disease. Alzheimer's The plot depicts two different gene expressions, with the x-axis representing sample and the y-axis representing gene expression value, and the horizontal line representing the gene expression mean.

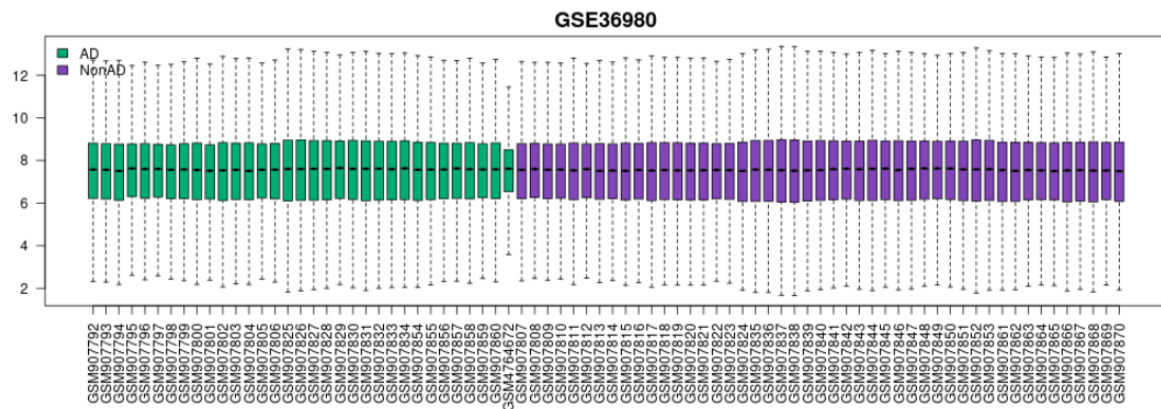


Fig:5.23 : Gene expression comparison between AD Vs Non-AD

5.12 . DENSITY PLOT

The density plot will show how data is distributed in continuous intervals. There are two types of coloured curves here: one that is blue and represents Non AD, and the other that is green and represents AD. The density is represented by the y-axis, while the intensity is represented by the x-axis. The higher the value, the higher the rate of expression.

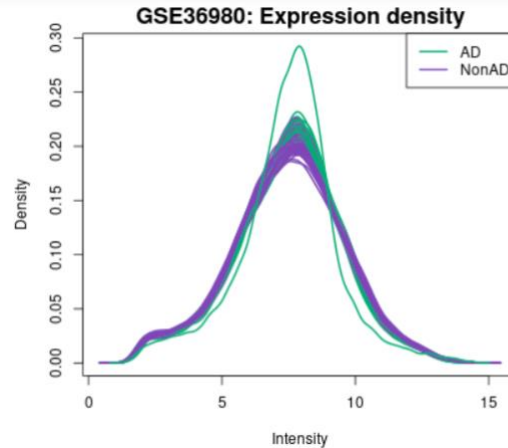


Fig:5.24 : Expression rate between AD Vs Non-AD

5.13 HISTOGRAM

When analysing the results, this is used to see the p-value distribution. The number of probes is on the y-axis, and the P adjusted value is on the x-axis.

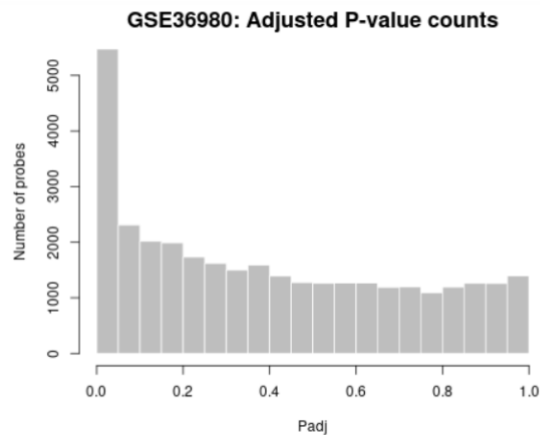


Fig: 5.25 : P-value distribution

5.14 MEAN VARIANCE

It examines the expression data's mean-variance relationship. This data can be used to determine variation. The plot's points represent genes, with each point representing a gene. The mean variance is represented by a red line, while the constant variance is represented by a blue line with no changes and the changes are held in the red line.

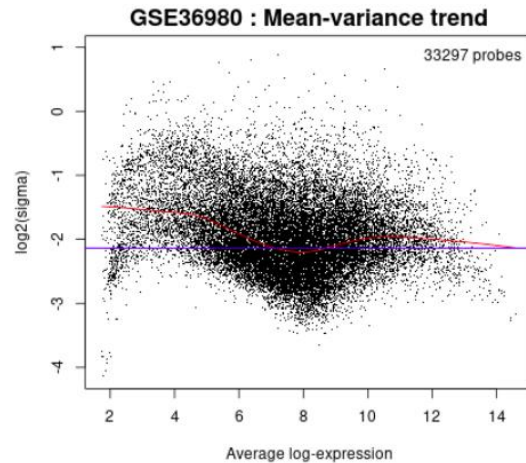


Fig: 5.26 : Mean-Variance

5.15 POPULATION GENETICS

For each variant, the population frequency of SNPs can be determined. Using Genome Asia 100k, it displays the allele count across the population with different statistics varying from variant to variant.

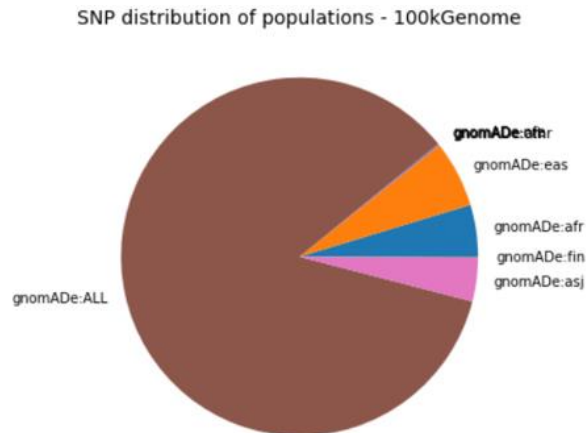


Fig: 5.27 : SNP distribution of population – 100k genome

5.16 GWAS DATA

The following is a list of SNPs discovered through GWAS studies. This explains the statistical significance of the SNPs. A total of 60 SNPs have been discovered.

Table 5.3 : SNPs identification through GWAS

CHROMOSOME	POSITION	MARKER NAME	EFFECT_ALLELE		NON_EFFECT_ALLELE		BETA	SE	PVALUE
	1	2256288	rs76805690	A	C	0.0389	0.0175	0.025980	
1	1	2261983	rs75379543	A	C	0.0427	0.0167	0.010780	
2	1	2263666	rs75273719	A	G	0.0502	0.0171	0.003361	
3	1	2263888	rs903904	T	C	0.0413	0.0169	0.014870	
4	1	2265070	rs80107522	C	T	0.0446	0.0168	0.007704	
5	1	4867850	rs12410445	G	T	-0.0050	0.0332	0.879100	
6	1	6046264	rs12142270	A	G	-0.0014	0.0416	0.972700	
7	1	6520312	rs3138158	T	C	-0.0522	0.0335	0.118600	
8	1	12401868	rs28551666	A	C	-0.0627	0.0179	0.000444	
9	1	14699523	rs10754853	C	T	0.0331	0.0141	0.019080	
10	1	14815696	rs74226815	G	T	0.0423	0.0212	0.046190	
11	1	15719953	rs11802154	A	G	0.2363	0.0590	0.000061	
12	1	15722130	rs59519309	T	G	0.2412	0.0602	0.000062	
13	1	15728786	rs7547778	T	A	0.2352	0.0547	0.000017	
14	1	15729378	rs111853844	A	T	0.2371	0.0593	0.000063	
15	1	15741716	rs78356043	G	A	0.2795	0.0659	0.000022	

16	1	15741823	rs113083027	G	A	0.2751	0.0662	0.000032
17	1	17357577	rs7550829	T	C	0.0663	0.0218	0.002292
18	1	22490581	rs112411925	T	C	0.1431	0.0728	0.049380
19	1	22493824	rs12405138	T	C	0.0534	0.0189	0.004748
20	1	26392824	rs35123100	A	C	0.0766	0.0290	0.008149
21	1	27397604	rs12410656	T	C	-0.0791	0.0349	0.023560
22	1	28281848	rs75443844	T	C	-0.1492	0.0570	0.008890
23	1	28287264	rs78664461	G	C	-0.1373	0.0522	0.008452
24	1	30051038	rs28716352	G	C	0.0629	0.0188	0.000810
25	1	30051885	rs61765975	G	A	0.0618	0.0187	0.000933
26	1	30054456	rs28535875	A	G	0.0621	0.0186	0.000823
27	1	30057494	rs28698112	C	T	0.0618	0.0185	0.000846
28	1	30058768	rs28548748	A	C	0.0650	0.0185	0.000456
29	1	30059295	rs28691898	G	A	0.0642	0.0185	0.000505
...	...	...	...	...	...	...	...	...
11602	22	41622419	rs5758268	T	A	0.0619	0.0143	0.000016
11603	22	41627775	rs5751069	G	C	0.0618	0.0143	0.000016
11604	22	41627924	rs9611522	T	C	0.0670	0.0150	0.000008

11605	22	42389500	rs2097562	C	T	-0.0544	0.0140	0.000102
11606	22	42389948	rs739296	A	G	-0.0547	0.0140	0.000093
11607	22	42390209	rs7290907	A	G	-0.0524	0.0140	0.000180
11608	22	42390225	rs7290655	G	C	-0.0540	0.0140	0.000114
11609	22	42396371	rs8138080	A	G	-0.0531	0.0140	0.000147
11610	22	42398195	rs11090066	C	T	-0.0539	0.0140	0.000119
11611	22	42401514	rs9611700	T	C	-0.0548	0.0141	0.000104
11612	22	42402399	rs2413663	A	T	-0.0562	0.0141	0.000069
11613	22	42403980	rs59993856	T	C	-0.0551	0.0140	0.000084
11614	22	42405657	rs9620007	G	C	-0.0558	0.0142	0.000090
11615	22	42410695	rs9611711	A	G	-0.0543	0.0140	0.000112
11616	22	42410993	rs9611712	C	T	-0.0554	0.0141	0.000083
11617	22	42418110	rs9607869	A	T	-0.0548	0.0140	0.000093
11618	22	44744127	rs138340382	T	C	0.2187	0.0815	0.007294
11619	22	44880343	rs13054771	T	G	0.1162	0.0343	0.000692
11620	22	45007914	rs8135966	G	A	0.0331	0.0153	0.029990
11621	22	45286318	rs12157651	C	A	0.1063	0.0366	0.003657
11622	22	45313654	rs75283360	A	G	0.1153	0.0366	0.001635

11623	22	45315254	rs7284925	G	T	0.1182	0.0367	0.001284
11624	22	46946127	rs738244	T	C	0.0356	0.0154	0.020520
11625	22	48209209	rs2208121	C	T	-0.0519	0.0151	0.000567
11626	22	48209434	rs1884697	T	C	-0.0519	0.0150	0.000554
11627	22	48209939	rs6008402	A	G	-0.0538	0.0153	0.000444
11628	22	48210968	rs74876332	C	T	-0.0538	0.0151	0.000377
11629	22	48682578	rs11090825	G	A	0.0916	0.0276	0.000913
11630	22	48684220	rs73169417	C	A	0.0302	0.0162	0.062720
11631	22	49215182	rs9628575	C	A	-0.0427	0.0173	0.013620

5.17 MANHATTAN PLOT

This plot is used to display the SNPs in the GWAS. The threshold is indicated by the horizontal line. In the X-axis, genomic coordinates are plotted, and in the Y-axis, negative P-values are plotted. Each dot represents a single nucleotide polymorphism (SNP). From this, expression levels can be observed.

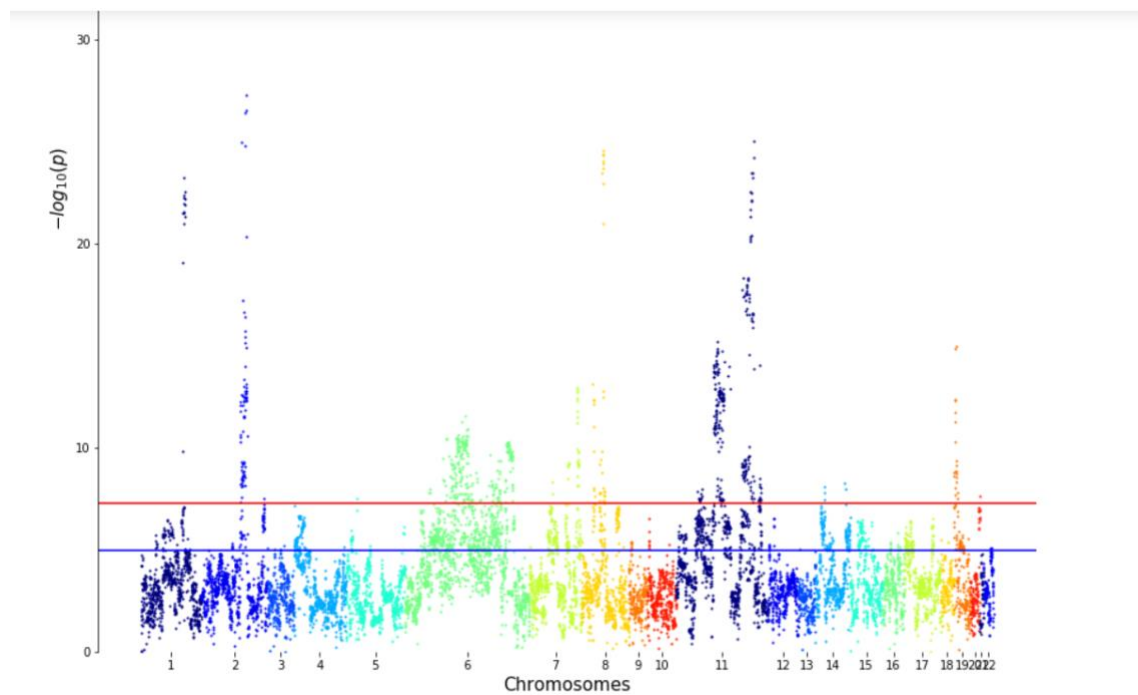


Fig:5.28. SNP distribution in Manhattan Plot

CHAPTER 6

SUMMARY AND CONCLUSION

6.1 SUMMARY

In post-mortem brain tissue from controls and Alzheimer's disease patients, there are a total of 46 differentially expressed genes. With $P_{adj}0.05$, these genes have a range of expression levels. The genes expressed in the up-regulated genes of AD Vs NonAd are expressed statistically significant with $P\text{-value}<0.5$ and the genes present in the outliers are expressed very differently with $P\text{-value}>0.5$. In the DEG of Alzheimer's disease, a total of 45 novel genes have been discovered, as well as false positives. There are a total of 82 SNPs that are shared between the Phenotype and Genotype of AD that was discovered using PenGenI and IGAP. A total of 60 SNPs have been identified between the SNPs of sample genome of AD and the IGAP through GWAS studies.

6.2 CONCLUSION AND FUTURE WORK

Data sets were obtained from patients with Alzheimer's disease, and different methods were used to detect SNPs in the genome of Braak stage III & Braak stage VI AD brain cells. SNPs that share the common genotype phenotype are identified and analysed using IGAP and PenGenI. Probability of each SNPs identified was a GWAS forecast. Population distribution with different allele frequency was observed with the SNPs and Genome Asia 100k for each SNP. DEGs were examined to determine which genes are up-regulated and which are down-regulated between non-AD AD Vs samples. Research is conducted between diseased and non-diseased gene expression analysis to obtain differentially expressed genes to distinguish between novel genes and false positives. Further research can be done on network analysis and the construction of a pipeline between Genome and AD Transcriptome.

REFERENCES

1. Du, Xiaoguang, Xinyi Wang, and MeiyuGeng. "Alzheimer's disease hypothesis and related therapies." *Translational neurodegeneration* 7.1 (2018): 1-7.
2. Wildburger, Norelle C., et al. "Diversity of amyloid-beta proteoforms in the Alzheimer's disease brain." *Scientific reports* 7.1 (2017): 1-9.
3. Sarlus, Heela, and Michael T. Heneka. "Microglia in Alzheimer's disease." *The Journal of clinical investigation* 127.9 (2017): 3240-3249.
4. Bellenguez, Céline, Benjamin Grenier-Boley, and Jean-Charles Lambert. "Genetics of Alzheimer's disease: Where we are, and where we are going." *Current opinion in neurobiology* 61 (2020): 40-48.
5. Bagyinszky, Eva, Vo Van Giau, and SeongSoo A. An. "Transcriptomics in Alzheimer's Disease: Aspects and Challenges." *International Journal of Molecular Sciences* 21.10 (2020): 3517.
6. Okasha, Samir. "Population genetics." (2006).
7. Hoie, Eric. "Alzheimer's disease: current treatments and potential new agents." *US Pharm* 44.1 (2019): 20-23.
8. Freudenberg-Hua, Yun, Wentian Li, and Peter Davies. "The role of genetics in advancing precision medicine for Alzheimer's disease—a narrative review." *Frontiers in medicine* 5 (2018): 108.
9. Bekris, Lynn M et al. "Genetics of Alzheimer disease." *Journal of geriatric psychiatry and neurology* vol. 23,4 (2010): 213-27.
10. Jansen, Iris E., et al. "Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk." *Nature genetics* 51.3 (2019): 404-413.
11. Hu, Xiaolan, et al. "Meta-analysis for genome-wide association study identifies multiple variants at the BIN1 locus associated with late-onset Alzheimer's disease." *PloS one* 6.2 (2011).
12. Hardy, John, and Valentina Escott-Price. "Genes, pathways and risk prediction in Alzheimer's disease." *Human molecular genetics* 28.R2 (2019): R235-R240.

13. Bagyinszky, Eva, Vo Van Giau, and SeongSoo A. An. "Transcriptomics in Alzheimer's Disease: Aspects and Challenges." *International Journal of Molecular Sciences* 21.10 (2020): 3517.
14. Penney, Jay, William T. Ralvenius, and Li-Huei Tsai. "Modeling Alzheimer's disease with iPSC-derived brain cells." *Molecular psychiatry* 25.1 (2020): 148-167.
15. Han, Xiaojuan, et al. "Astrocyte senescence and Alzheimer's disease: a review." *Frontiers in aging neuroscience* 12 (2020).
16. Dafsari, Forugh S., and Frank Jessen. "Depression—an underrecognized target for prevention of dementia in Alzheimer's disease." *Translational Psychiatry* 10.1 (2020): 1-13.
17. Kunkle, Brian W., et al. "Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates A β , tau, immunity and lipid processing." *Nature genetics* 51.3 (2019): 414-430.
18. Nho, Kwangsik, et al. "Genome-wide transcriptome analysis identifies novel dysregulated genes implicated in Alzheimer's pathology." *Alzheimer's & Dementia* 16.9 (2020): 1213-1223.
19. Yang, Yu, et al. "Genetic and expression analysis of COPI genes and Alzheimer's disease susceptibility." *Frontiers in genetics* 10 (2019): 866.
20. Hao, Sicheng, et al. "Prediction of Alzheimer's disease-associated genes by integration of GWAS summary data and expression data." *Frontiers in genetics* 9 (2019): 653.
21. Angelucci, Francesco, et al. "MicroRNAs in Alzheimer's disease: diagnostic markers or therapeutic agents?" *Frontiers in pharmacology* 10 (2019): 665.
22. Hu, Yang, et al. "rs1990622 variant associates with Alzheimer's disease and regulates TMEM106B expression in human brain tissues." *BMC medicine* 19.1 (2021): 1-10.
23. Morabito, Samuel, et al. "Integrative genomics approach identifies conserved transcriptomic networks in Alzheimer's disease." *Human molecular genetics* 29.17 (2020): 2899-2919.

24. Verheijen, Jan, and Kristel Sleegers. "Understanding Alzheimer disease at the interface between genetics and transcriptomics." *Trends in Genetics* 34.6 (2018): 434-447.
25. Brown, Joseph, Meg Pirrung, and Lee Ann McCue. "FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool." *Bioinformatics* 33.19 (2017): 3137-3139.
26. Langmead, Ben, et al. "Scaling read aligners to hundreds of threads on general-purpose processors." *Bioinformatics* 35.3 (2019): 421-432.
27. Langmead, Ben, and Steven L. Salzberg. "Fast gapped-read alignment with Bowtie 2." *Nature methods* 9.4 (2012): 357.
28. Li, Heng, et al. "The sequence alignment/map format and SAMtools." *Bioinformatics* 25.16 (2009): 2078-2079.
29. Yao, Zhen, et al. "Evaluation of variant calling tools for large plant genome re-sequencing." *BMC bioinformatics* 21.1 (2020): 1-16.
30. GenomeAsia100K Consortium. "The GenomeAsia 100K Project enables genetic discoveries across Asia." *Nature* 576.7785 (2019): 106.