**ISTANBUL TECHNICAL UNIVERSITY ★ FACULTY OF SCIENCE AND LETTERS**

# THE STUDY OF RNA LEVELS TO UNDERSTAND GENE EXPRESSION PROFILE CHANGES IN ALZHEIMER'S DISEASE

**B.Sc. THESIS**

**Fatma Betül BOZKURT**

**Department of Molecular Biology and Genetics**

**JULY 2020**

**ISTANBUL TECHNICAL UNIVERSITY ★ FACULTY OF SCIENCE AND LETTERS**

# THE STUDY OF RNA LEVELS TO UNDERSTAND GENE EXPRESSION PROFILE CHANGES IN ALZHEIMER'S DISEASE

**B.Sc. THESIS**

**Fatma Betül BOZKURT**
**(090140049)**

**Department of Molecular Biology and Genetics**

**Thesis Advisor: Assoc. Prof. Dr. Halil Bayraktar**

**JULY 2020**

**İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN EDEBİYAT FAKÜLTESİ**

# ALZHEIMER HASTALIĞINDA GEN İFADE PROFİLİ DEĞİŞİKLİKLERİNİ ANLAMAK İÇİN RNA DÜZEYLERİNİN İNCELENMESİ

**LİSANS TEZİ**

**Fatma Betül BOZKURT**
**(090140049)**

**Moleküler Biyoloji ve Genetik Bölümü**

**Tez Danışmanı: Doç. Dr. Halil BAYRAKTAR**

**TEMMUZ 2020**

Fatma Betül Bozkurt, a B.Sc. student of ITU Faculty of Science and Letters student ID 090140049, successfully defended the thesis entitled “THESIS TITLE”, which she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

**Thesis Advisor :** **Assoc. Prof. Dr. Halil BAYRAKTAR** ..............................
Istanbul Technical University

**Jury Members :** **Assoc. Prof. Dr. Gizem DİNLER DOĞANAY** ..............................
Istanbul Technical University

**Assist. Prof. Dr. Bülent BALTA** ..............................
Istanbul Technical University

**Date of Submission : 20.07.2020**
**Date of Defense : 24.07.2020**

***To my family,***

## FOREWORD

This thesis was written for my Bachelor’s degree in Molecular Biology and Genetics at Istanbul Technical University. The main topic of the thesis is Alzheimer’s disease and its related gene expression profile changes.

I would like to express my gratitude to my thesis advisor Assoc. Prof. Dr. Halil BAYRAKTAR for his guidance, support, and recommendations. I am very grateful that he shared his valuable time and knowledge with me. Over the course of my graduation project, I learned a lot from him.

I am grateful for my husband’s love, support, and motivational talks throughout my graduation project. I would like to thank also my parents and siblings for always being a source of motivation and being there for me wherever and whenever I needed them.

July 2020 Fatma Betül BOZKURT

## TABLE OF CONTENTS

## ABBREVIATIONS

| | |
|---|---|
| **Aβ42** | **:** Amyloid- β |
| **AD** | **:** Alzheimer’s Disease |
| **APOE** | **:** Apolipoprotein E |
| **cDNA** | **:** Complementary DNA |
| **CDx** | **:** Companion Diagnostic |
| **circRNA** | **:** Circular RNA |
| **CSF** | **:** Cerebrospinal Fluid |
| **DNA** | **:** Deoxyribonucleic Acid |
| **EOAD** | **:** Early-Onset Alzheimer’s Disease |
| **GWAS** | **:** Genome-Wide Association Study |
| **HER2** | **:** Human Epidermal Growth Factor Receptor 2 |
| **lncRNA** | **:** Long non-coding RNA |
| **LOAD** | **:** Late-Onset Alzheimer’s Disease |
| **MCI** | **:** Mild Cognitive Impairment |
| **miRNA** | **:** Micro RNA |
| **mRNA** | **:** Messenger RNA |
| **ncRNA** | **:** Non-coding RNA |
| **NGS** | **:** Next-Generation Sequencing |
| **PCA** | **:** Principal Component Analysis |
| **piRNA** | **:** Piwi-activating RNA |
| **P-tau** | **:** Phosphorylated tau |
| **qPCR** | **:** Quantitative Polymerase Chain Reaction |
| **RNA-Seq** | **:** RNA Sequencing |
| **RNA** | **:** Ribonucleic Acid |
| **rRNA** | **:** Ribosomal RNA |
| **siRNA** | **:** Small interfering RNA |
| **snoRNA** | **:** Small nucleolar RNA |
| **SNP** | **:** Single Nucleotide Polymorphism |
| **tRNA** | **:** Transfer RNA |
| **T-tau** | **:** Total tau |

## LIST OF TABLES

## LIST OF FIGURES

# THE STUDY OF RNA LEVELS TO UNDERSTAND GENE EXPRESSION PROFILE CHANGES IN ALZHEIMER’S DISEASE

## SUMMARY

The genetics of Alzheimer’s disease is highly complex, and numerous risk genes have been identified so far. Transcriptome studies are widely used for understanding the genetics and pathogenesis of Alzheimer’s disease, and to discover novel biomarkers for the disease. This thesis mainly aims to demonstrate the alterations observed in gene expression profiles of Alzheimer’s disease patients by analyzing their RNA levels, to identify genes linked to the disease development, and investigate the involvement of these genes in pathways linked to the disease pathology. For these purposes, the RNA-Seq dataset containing the total RNA isolated from frozen brain tissue samples (lateral temporal lobe) from 8 young healthy individuals, 10 old healthy individuals, and 12 aged diseased individuals was analyzed by using Matlab. The examination of mean values revealed two genes (MECR and PTRPRD-AS2) having distinct RNA levels in young healthy brains, and three genes (LOC283440, PRKACB, and LINC01372) having distinct RNA levels in aged diseased brains. Two sample t-test was applied to healthy and diseased individuals’ normalized RNA levels to determine the most significant genes for detecting Alzheimer’s disease. 80 genes were reported to have significantly different RNA levels in healthy and diseased individuals, and proposed as potential biomarkers, including also the LOC283440, PRKACB, and LINC01372 genes. It is suggested that, apart from MECR and PRKACB genes whose association with Alzheimer’s disease have already been reported, the identified genes in this study might be possibly involved in pathways related to Alzheimer’s disease pathology. Furthermore, the analysis of correlation coefficients detected 4 gene pairs (SFR1 & CPPED1, CPPED1 & SGK1, DOCK5 & ATP10B, SFR1 & SGK1) having strongly correlated RNA levels in all samples, and two AD patients (21st and 25th individuals of the data set) having highly correlated RNA levels. The principal component analysis and cluster analysis were performed to observe different groups for healthy and diseased individuals. It was found that healthy young and healthy old individuals were located in the same group, whereas aged diseased individuals formed another group with some exceptions (13th, 18th, 20th, 27th, and 30th individuals of the data set). It is suggested that these individuals might have been misdiagnosed. Lastly, binary logistic models were created to be able to predict the risk of Alzheimer’s disease for an individual based on RNA intensities of priorly determined significant genes. Two statistically significant models were obtained by using the RNA intensity values of first three and first seven genes of those 80 important genes. It is concluded that gene expression profiles of individuals with Alzheimer’s disease are significantly altered, and transcriptomics provide a significant insight to the genetics and pathology of the disease.

# ALZHEIMER HASTALIĞINDA GEN İFADE PROFİLİ DEĞİŞİKLİKLERİNİ ANLAMAK İÇİN RNA DÜZEYLERİNİN İNCELENMESİ

## ÖZET

Alzheimer hastalığının genetiği oldukça karmaşıktır ve şimdiye kadar çok sayıda risk geni tanımlanmıştır. Transkriptom çalışmaları, Alzheimer hastalığının genetik ve patogenezini anlamak ve hastalık için yeni biyobelirteçler keşfetmek amacıyla yaygın olarak kullanılmaktadır. Bu tezin asıl amacı Alzheimer hastalarının gen ekspresyon profillerinde gözlemlenen değişiklikleri RNA düzeylerini analiz ederek anlamak, hastalığın gelişimine ilişkin genleri tanımlamak, ve bu genlerin hastalık patolojisiyle ilişkili olan yolaklara katılımlarını araştırmaktır. Bu amaçla, 8 genç sağlıklı bireyin, 10 yaşlı sağlıklı bireyin ve 12 yaşlı hastalıklı bireyin dondurulmuş beyin dokusu örneklerinden (lateral temporal lob) izole edilen toplam RNA düzeylerini içeren RNA-Seq veri kümesi Matlab kullanılarak analiz edilmiştir. RNA düzeylerinin ortalama değerlerin incelenmesi ile, genç sağlıklı bireylerde farklı RNA düzeylerine sahip iki gen (MECR ve PTRPRD-AS2) ve yaşlı hastalıklı bireylerde farklı RNA düzeylerine sahip üç gen (LOC283440, PRKACB ve LINC01372) tespit edilmiştir. Alzheimer hastalığının tespiti için en önemli olan genleri belirlemek üzere sağlıklı ve hastalıklı bireylerin normalize edilmiş RNA seviyelerine iki örnek t testi uygulanmıştır. 80 genin sağlıklı ve hastalıklı bireylerde önemli ölçüde farklı RNA seviyelerine sahip olduğu saptanmış ve bu genlerin (LOC283440, PRKACB ve LINC01372 genleri de dahil olmak üzere) potansiyel biyobelirteçler olabileceği önerilmiştir. Alzheimer hastalığı ile ilişkisi daha önce yapılan çalışmalarda belirtilmiş olan MECR ve PRKACB genlerinin yanı sıra, bu çalışmada tanımlanan genlerin Alzheimer hastalığı patolojisine ilişkin yolaklarda yer alıyor olabileceği belirtilmiştir. Ek olarak, korelasyon katsayılarının analizi ile tüm örneklerde güçlü korelasyona sahip 4 gen çifti (SFR1 & CPPED1, CPPED1 & SGK1, DOCK5 & ATP10B, SFR1 ve SGK1) ve çok benzer RNA düzeylerine sahip olan iki AD hastası (veri setinin 21. ve 25. bireyleri) tespit edilmiştir. Sağlıklı ve hastalıklı bireylerin farklı gruplar oluşturduklarını gözlemlemek amacıyla temel bileşenler analizi ve kümeleme analizi yapılmıştır. Bazı istisnalar dışında (veri kümesinde bulunan 13, 18, 20, 27, ve 30. bireyler), sağlıklı genç ve sağlıklı yaşlı bireylerin aynı grupta yer aldığı, yaşlı hastalıklı bireylerin ise farklı bir grup oluşturdukları bulunmuştur. Bulunması gereken gruptan farklı grupta yer alan bireylere yanlış teşhis koyulmuş olabileceği belirtilmiştir. Son olarak, bir bireyin RNA düzeylerine bakarak Alzheimer hastalığı riskini tahmin edebilmek amacıyla ikili lojistik regresyon modelleri oluşturulmuştur. Bu çalışmada önemli olduğu saptanmış olan 80 genin ilk üç ve ilk yedi geninin RNA düzeyleri kullanılarak oluşturulan ikili lojistik regresyon modellerinin istatistiksel olarak anlamlı olduğu ve risk tespitinde kullanılabileceği belirlenmiştir. Sonuç olarak, Alzheimer hastalığına sahip olan bireylerin gen ekspresyon profillerinin önemli ölçüde değiştiği ve transkriptom analizininin hastalığın genetiğinin ve patolojisinin anlaşılmasına önemli katkılar sağladığı görülmüştür.

# 1 INTRODUCTION

## 1.1 Alzheimer's Disease

Alzheimer's disease (AD) is a common health problem which is listed in the top ten causes of global deaths. Even though the deaths caused by other major health issues have been decreasing, there has been a dramatic increase in the number of people dying from AD ("2019 Alzheimer's disease facts and figures," 2019; Lane et al., 2018; World Health Organization, 2018). AD is classified among age-related chronic diseases and it is an irreversible neurodegenerative disorder. In elderly patients, the appearance of the disease has observed as the dominating form of senility having an impact on ten percent of individuals above the age of 65, and half of the people above the age of 85. Even though the specific reasons for the development of the disease are still unclear, it is suggested by a considerable number of genetic and pathological studies carried out over the last couple of decades that the process of the disease is characterized by the deposition of β-amyloid plaques and neurofibrillary hyperphosphorylated tau tangles. This deposition results in loss of contacts between neurons, ultimately leading to the neural cell death. However, at the initial stages of AD, the correlation between amyloidosis and the degree of deterioration of cognitive functions is low, indicating that there might be other factors promoting the development of the disease. (Alkadhi & Eriksen, 2011; National Institutes of Health, 2019a; Weller & Budson, 2018). It has been found that there are many different target proteins contributing to the development of AD; therefore, it is regarded as a multifactorial disease (Carreiras et al., 2013). There are two types of Alzheimer's disease which are called as early-onset AD (EOAD) and late-onset AD (LOAD); and for both types, genetics plays a significant role. Most of the AD patients has the late-onset type where the symptoms start to appear at the age of 65 and later. Variation in the apolipoprotein E (APOE) gene on chromosome 19 was found to be a risk factor for developing the disease. The role of the APOE gene is the involvement in the production of a protein which assists in carrying cholesterol and different types of fats in the bloodstream. There are various forms of APOE. One of them is APOE ε2 and it is rare when compared to other alleles. It might be involved in protection against AD. Usually, patients with this allele develops AD at older ages contrary to patients having the APOE ε4 gene. The most common form is APOE ε3, and its role is considered as neutral for the development of the disease. The last allele is APOE ε4, and it rises the risk of developing Alzheimer's disease; additionally, it is related to early-onset of the disease. Thus, APOE ε4 gene is regarded as a risk-factor. Nevertheless, having an APOE ε4 allele does not cause that person to develop AD precisely. Even though an individual carry an APOE ε4 allele, he/she may not develop the disease; also, it is possible that AD patients may not have any of these alleles (National Institutes of Health, 2019a). There have been four main approaches in the AD genetics studies which are genetic linkage analysis, genome-wide association studies (GWASs), investigation of candidate genes, and next generation sequencing technology. The first landmark to figure out the genetic basis of AD research for families exhibiting autosomal dominant inheritance was the genetic linkage analysis. The purpose of these studies is to determine the chromosomal regions related to disease, but not to determine a particular gene or a mutation linked to a disease. By means of genetic linkage studies conducted in families, dominantly inherited mutations in amyloid-β precursor protein (APP) on chromosome 21q, presenilin 1 (PSEN1) on 14q, and presenilin 2 (PSEN2) on 1q have been found to be associated with early-onset of AD. The ε4 allele of the APOE is referred as the only genetic risk factor found for late-onset of the disease, and it was also determined by genetic linkage studies. Linkage mapping is very beneficial for identification of monogenic characteristics in EOAD; however, it has been a failure for determination of risk factors in LOAD, presumably owing to the complex characteristics with unknown variants (Giri et al., 2016; Tanzi & Bertram, 2005; Verghese et al., 2011). Genetics studies have been revolutionized by the microarray technology which is enabling the simultaneous analysis of millions of single-nucleotide polymorphisms (SNPs) in a wide array of samples. In addition to APOE, more than 20 loci related to an enhanced susceptibility for LOAD were identified by GWASs. Even though several susceptibility genes have been successfully determined by GWASs, detection of rare variants is not possible by these studies (Giri et al., 2016; Karch et al., 2014). In studies of candidate genes, comparison of genetic variations of individuals with a certain disease and healthy individuals are made. Significancy of difference between the frequencies of case and control is evaluated to unveil susceptibility genes. Candidate gene studies revealed that APOE alleles are strongly involved in late-onset AD. More than thousand of candidate genes were examined for susceptibility of AD, usually having incoherent outcomes. Current knowledge on disease pathology is crucial for the success of candidate gene studies (Giri et al., 2016; Hattersley & McCarthy, 2005). NGS technologies offer rapid and cost-effective sequencing approaches allowing the sequencing of an entire genome in less than a day. By the use of NGS technology, the comprehension of several Mendelian neurological disorders and complex neurological diseases has become possible; since, the rare variants of disease can be determined, and small mutations can be uncovered by the NGS technology. In recent studies, NGS technology has been used for the identification of genetic factors in small families having EOAD with unknown causes. The determination of rare susceptibility altering alleles in APP, triggering receptor expressed on myeloid cells 2 (TREM2), and phospholipase D family member 3 (PLD3) has been a major success of NGS approach in AD (Bertram, 2016; Giri et al., 2016). As stated before, three genes have been found to be associated with early-onset AD are APP on chromosome 21q, PSEN1 on 14q, and PSEN2 on 1q. More than two hundred penetrant mutations have been detected in these three genes resulting in inherited AD. The APP gene's location is chromosome 21q21.3 and it functions in the development of the nervous system, synaptogenesis and repair of synaptic connections, and amyloid beta production. PSEN1 gene is located on chromosome 14q24.3. It functions in γ-Secretase activity, intracellular signal transduction, and amyloid beta production. PSEN2 gene is located on 1q42.13. It also functions in γ-Secretase activity and amyloid beta production. Besides, it functions in synaptic plasticity. These three genes are involved in the APP processing pathway. In the pathogenesis of AD, modified APP processing and Aβ accumulation are very significant. Since several genes and environmental factors are possibly involved in the development of LOAD, the complexity of the genetics of LOAD is greater when compared to the genetic complexity of EOAD. In most of the cases, LOAD is sporadic, having no family history of AD. A clustering within three pathways (shown in Figure 1) has been observed for most of the genes detected by GWASs that might be associated with the amyloid cascade or tau pathology (Giri et al., 2016).

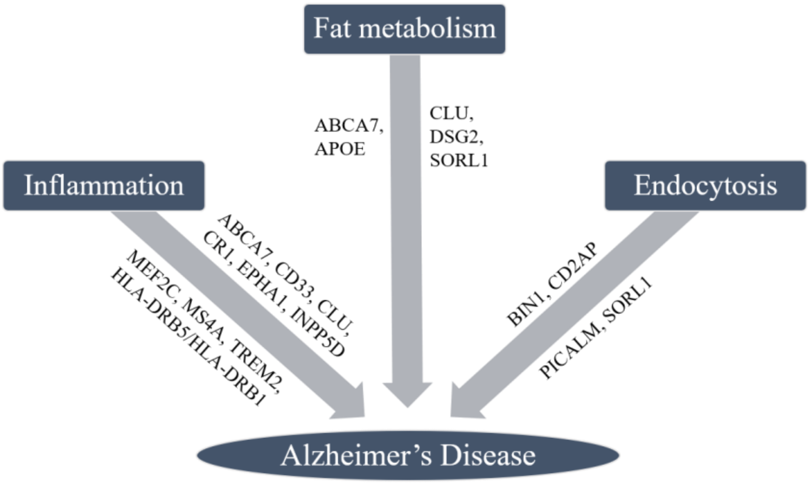

**Figure 1.1.1** Major pathways involved in the development of AD and genes associated with these pathways

APOE, SORL1, ABCA7, and CLU genes are involved in cholesterol metabolism. It has been found that having high cholesterol levels in midde ages resulted in the increased risk of developing Alzheimer's disease in older ages. Various GWASs have revealed that the immune system is involved in the LOAD pathology. One of the pathological characteristics of AD is neuroinflammation. CR1, CD33, MS4A, ABCA7, EPHA1, TREM2, and CLU are the genes which are involved in development of LOAD. For neurotransmission and neural damage response, endocytosis process is very crucial. Several genes involved in the regulation of endocytosis have been revealed by GWASs in LOAD: BIN1, CD2AP, PICALM, EPHA1, SORL1, etc. APP trafficking has a very significant role in the pathogenesis of AD, and the majority of these genes are involved in this process. Additional genes involved in LOAD are: HLA-DRB5/HLA-DRB1, INPP5D, and MEF2C (involved in immune and inflammatory response); CASS4, CELF1, and NME8 (involved in axoplasmis transport and cytoskeletal function); PTK2B (implicated in hippocampal synaptic function); ZCWPW1 (epigenetic regulator); FERMT2 (involved in angiogenesis and tau pathology); and SLC24H4-RIN3 having a likely cardiovascular risk. Most of their functions have not been identified yet, or poorly defined. Their involvement in various different pathways is possible. The common variants (but not the rare variants) linked to LOAD have been successfully determined by GWASs. The determination of novel genes having low-frequency variants is possible due to the improvements in sequencing technologies including robust techniques such as whole exome sequencing and whole genome sequencing. By the use of these techniques, ADAM10, AKAP9, UNC5C, PLD3, and TREM2 genes have been found to be related with AD (Giri et al., 2016).

### 1.2 Biomarkers and Personalized Medicine

Personalized or precision medicine is a therapeutic and preventive approach which considers the variations in the genome, environment, and lifestyles of individuals. It provides personalization on the basis of factors which might have an effect on treatment response. The significance of biomarkers is on the increase, especially for personalized medicine which comprises several applications such as diagnosis, prognosis, and targeted therapy selection. A biomarker is basically a molecule showing a change in physiology from the normal physiology. For instance, any particular change on DNA, RNA, or protein level in a cancer cell can be regarded as a molecular biomarker. It can be defined as a trait which is a measure of normal biological and pathogenical events, or pharmacological response to a therapy. Biomarkers are fundamental tools for matching the patients with proposed treatments using specific drugs, and allow for ‘providing the right treatment to the right patient, at the right dose at the right time'. Therefore, the discovery and qualification of clinically effective novel biomarkers is crucial for increasing the number of conditions where the usage of personalized approach is possible. Substantial technologies for molecular diagnostic have been used for identification of biomarkers of several diseases including cancer, metabolic disorders, infectious diseases, and neurological disorders. Two categories of novel molecular diagnostics related to personalized medicine are pharmacogenetic tests and pharmacogenomic tests. A pharmacogenetic test is used for studying the variations in DNA sequences associated with the absorption and disposition of the drug or the action of the drug between individuals, whereas A pharmacogenomic test is used for analyzing the variations in whole-genome or candidate gene, single nucleotide polymorphisms, haplotype markers, changes in expression of genes, or deactivation that might be associated with a pharmacological function and a response to therapy among individuals. Sometimes, the biomarker can be an alteration in a pattern or a profile, as opposed to being a change in a particular marker. By the use of diagnostic systems including DNA microarrays and proteomics, multiple biomarkers can be simultaneously evaluated. Biomarkers might be categorized according to their biological aspects, measurement, and aim of use. In relation to biological aspects, various characteristics such as morphology, physiology, and molecular features can be used. The objective read-out of a biomarker is the second significant feature. It implies that the obtained outcome from a biomarker is not observer-dependent. With regard to understanding the results, a biomarker test might have a qualitative, a semi-quantitative, or a quantitative outcome. The aim of use of the biomarker test is dependent upon the result that would like to be obtained. The determination of the biomarker type is made by their particular application such as diagnosis, prognosis, prediction, safety evaluation, disease monitoring, analysis of drug effects, and as a surrogate marker substituting for a clinically meaningful endpoint in clinical trials. In personalized medicine, prognostic and prediction biomarkers have a major role. Apart from the therapy, the probability of disease results can be decided by the use of prognosis biomarkers which might be affecting the decision-making for treatments. Predictive biomarkers are essential for efficacy and safety prediction. The efficacy prediction is to detect people from a patient population which will probably take the advantage of a certain therapy, while the safety prediction is to determine individuals which might exhibit a toxic response. General prediction markers for particular treatment approaches such as anti-hormone therapy for breast cancer might be distinguished from companion diagnostic (CDx) biomarkers. CDx are developed together with a particular drug, and on the drug label it is usually stated that their application is obligatory. For instance, tests measuring the expression of estrogen receptors in breast cancer are general prediction markers, while approved tests measuring the expression of HER2 like DAKO Herceptest® for Herceptin® (trastuzumab) treatment are companion diagnostic biomarkers. In personalized medicine, other class of markers are safety prediction markers which are applied before the therapy for selecting patients. They differ from toxicology biomarkers applied to observe and identify toxic events. Even though the aim of use of prognostic and predictive markers are differ from each other, some biomarkers can be used for both purposes. For example, subtyping in breast cancer is applicable for both prognosis and stratification of patients. It has been found that the vulnerability to late-onset AD has been increased by various genetic and epigenetic risk factors; thus, integration of those factors into the strategy of Alzheimer's disease prevention for peculiar remedies in a targeted manner might be profitable (Berkowitz et al., 2018; Chaffey & Silmon, 2016; Jain, 2015; Landeck et al., 2016; Ziegler et al., 2012). It is supported by several clinical studies that the essential components of pathophysiology of AD is reflected by the most important Alzheimer's disease cerebrospinal fluid (CSF) biomarkers amyloid-β (Aβ42), total tau (T-tau), and phosphorylated tau (P-tau). Significantly, it is shown by a great number of medical research studies that these biomarkers give directly related information for diagnostics even in the early stages of the disease. Since there is heterogeneity in the pathology of late-onset AD, it is important for the AD CSF biomarker toolbox to be expanded with novel biomarkers (Blennow & Zetterberg, 2018).

## 1.3 RNA Sequencing (RNA-Seq) Analysis

The process of the transcription of the genetic information stored in deoxyribonucleic acid (DNA) into ribonucleic acid (RNA), and then its translation into proteins is explained by the central dogma of molecular biology. The phenotype of an organism is defined by the final gene expression which is affected by also environmental conditions. The identification of the cell and the regulation of its biological activities are determined by the transcription of a group of genes into RNA molecules which are wholly called as transcriptome. These RNA molecules are crucial for the comprehension of functional DNA elements and the development of diseases. The transcriptome is remarkably complex and it is comprised of coding and non-coding types of RNA. Throughout history, studies were concentrated mostly on messenger RNAs (mRNAs) since the proteins are encoded by the genetic information. Apart from mRNA, there are several types of functional non-coding RNA (ncRNA) molecules such as ribosomal RNAs (rRNA), transfer RNAs (tRNA), small nucleolar RNAs (snoRNAs) as being the most known types; and there are also novel types which are small non-coding RNAs, microRNA (miRNA), piwi-interacting RNAs (piRNAs), long non-coding RNAs (lncRNAs), small interfering RNAs (siRNAs), and circular RNAs (circRNAs) (Dana et al., 2017; Kukurba & Montgomery, 2015; Santer et al., 2019). Preliminary gene expression investigations are confined to measurement of only one form of RNA transcripts since they were based on low-throughput techniques such as quantitative polymerase chain reaction (qPCR) and northern blotting. Past couple of decades, new methods has developed to make the transcriptomics studies possible. Firstly, microarray technologies based on hybridization were used for conducting these studies. Then, sequence-based and tag-based techniques were developed. Though, none of these techniques could provide sensitive measurement of gene expression levels of splicing variants, and novel genes cannot be discovered by them. Besides, there are many other limitations to these methods such as the arduous work of cloning sequence tags, the great cost of automated Sanger sequencing, and the necessity of considerable amount of RNA input. The advancement of high-throughput next-generation sequencing (NGS) has vastly changed the discipline of transcriptomics studies by allowing the analysis of RNA via complementary DNA (cDNA) sequencing, which is called as RNA sequencing (RNA-Seq). A better way for the comprehension of the multifaceted nature of the transcriptome is provided by RNA-Seq technology, as it has prominent advantages when compared to former methods. By means of RNA-Seq, a more elaborate information of gene expression, differential splicing, and allele-specific expression can be obtained. Latest developments in the RNA-Seq process, from preparation of samples to bioinformatic data analysis, have provided profound transcriptome profiling for explication of diverse pathological and physiological circumstances. Presently, various different features of RNA biology such as translation, single-cell gene expression, and the structure of RNA can be studied by RNA-Seq techniques. Novel applications including spatial transcriptomics, as well as direct and long-read RNA-seq methods, and improved data analysis tools assist to completely comprehend the biology of RNA (Kukurba & Montgomery, 2015; Stark et al., 2019). RNA-seq have been widely used in several research on Alzheimer's disease. Bennett and Keeney (2018) have used RNA-seq for comparing gene expression in susceptible regions of Alzheimer's and Parkinson's disease brains. Annese et al. (2018), have used RNA-seq for profiling the whole transcriptome of patients with LOAD to provide understanding on the disease pathogenesis. Moreover, in their study Roy, Sarkar, Parida, Gosh, and Mallick (2017) have discovered piRNA dysregulations in AD and their potential roles in disease development by the use of small RNA sequencing.

## 1.4 Principal Component Analysis (PCA)

Principal component analysis has been used in several different areas such as biology, physics, and engineering etc. for decades. It mainly aims to reduce the dimensionality of a big data having multiple variables, and to determine fewer variables that are capable of summarizing the data (Bartholomew, 2010). It is an unsupervised method which requires no prior group information and is beneficial for studying possible classification of samples in a research. A set of principal components (PCs) are extracted by PCA in accordance with the distribution of the data set. The first principal component holds the highest variation, whereas the second principal component holds the second highest variation which is perpendicular to the first one. It is assumed that the distinction between the data points might be indicated prominently by the coordinates having higher variations, whereas the coordinates with minor variations might cause noise in the data and must be neglected (Liang, 2013; Roessner et al., 2011). There are several studies on Alzheimer's disease where PCA have been applied, the following are examples of these studies. Pagani, Salmaso, Rodriguez, Nardo, and Nobili (2009) have performed PCA to determine the distribution of regional cerebral blood flow between AD patients and controls. Campbell et al. (2013), have applied PCA to Pittsburgh compound B (PiB) PET imaging to analyze the similarity between in vivo amyloid beta (Aβ) pattern in Parkinson disease with intellectual disability and the pattern observed in Alzheimer's disease (AD). Moreover, Shigemizu et al. (2019), have used supervised PCA to build models for risk prediction for patients with dementia including AD by using miRNA expression data.

## 1.5 Cluster Analysis

Cluster analysis is a computer-based method which was developed by several research areas such as computer science, decision science, statistics, and pattern recognition. It is used in various fields including life science, social science, and geoscience (Lee, 1981). Cluster analysis is performed for grouping similar findings into coherent subgroups. These subsets might exhibit patterns associated with the case which the research is mainly focusing on. The comparison between samples and several clustering algorithms based upon various approaches might be evaluated by a distance function. Firstly, the similarity between observed values is measured. After these observed values start to form clusters, resemblance between clusters is calculated. For computing similarity, numerous measures are available such as Euclidean and Manhattan distance, correlation, and mutual information etc. Furthermore, it is likely to obtain diverse patterns of clustering by using different merging approaches. Since there are several options for the user, the outcome of clustering might be regarded as subjective. The conventional clustering approach only groups the variables or the observations distinctly. Alternatively, biclustering can be performed to group both variables and observations which might be used for biomarker studies. Hierarchical clustering is an insightful method to group data into clusters when there is no prior information about the count of the clusters. It uses a tree called dendrogram for organizing the data, leaves of this tree point out observations individually, whereas the association between clusters is shown by the branching. There are two different approaches available for hierarchical clustering: agglomerative and divisive. Practically, agglomerative approach is widely used. In this method, distances between observations and clusters are determined by a linkage function, and new clusters are created by grouping the nearest data points at every iteration. There are alternative methods to hierarchical clustering where a prior information about cluster counts is required. The K-means algorithm and the fuzzy-c means algorithm might be given as examples to those alternative ways of clustering (Boccard & Rudaz, 2013). The evaluation of Alzheimer's disease process and development is possible by clustering analysis which might display the pattern and foresee the results of the disease. Several studies have been conducted on AD where cluster analysis is used. These studies include the use of cluster analysis for displaying significant changes between female and male AD patients, for determination of pathological subtypes of AD, for identification of the effect of several variables on the occurrence of AD and mild cognitive impairment (MCI), etc. (Armstrong & Wood, 1994; Gamberger et al., 2016; Hamou et al., 2011).

## 1.6 Generalized Linear Models

Generalized linear models have been used in several fields including biology, biopharmaceuticals, engineering, quality control etc. (Myers et al., 2012). The distribution of a dependent variable (usually indicated as y) with regard to one ore several independent variables (usually indicated as x1, x2, and so on.) is defined by regression models. The ordinary linear regression is the most widely used among these models. In this model, the normal random variable y has a mean value which is a linear function of independent variables, b0 + b1*x1 + b2*x2 + …, and a constant variance. A straight line with normal distributions around every point might depict the simplest form of only one independent variable x. Generalized linear models are a wide group of models which could be listed as follows: linear regression, ANOVA, ANCOVA, logistic regression, loglinear, Poisson regression, and multinomial response. In generalized linear models, the average value for the dependent variable is formed by transforming a linear function of independent variables into a monotone non-linear function, g(b0 + b1*x1 + b2*x2 + …), where the reverse of g is referred to as the "link" function. Logit link and the log link might be given as examples to link funcions. Besides, the distribution of y does not have to be a normal distribution, it might also be a binomial, Poisson, or multinomial distribution. Logistic regression is a remarkable and powerful technique for predicting a dichotomous outcome (e.g., success/failure, diseased/healthy, etc.). In contrast to linear regression which predicts a numerical value, logistic regression predicts a binary value. Owing to the logistic link of the logistic regression model, the range of the expected proportions (probability) is restricted between 0 to 1. Logistic regression might be used to predict if an individual has a particular disorder or not. Linear regression cannot be used for this purpose, since there might be values lower than 0 or higher than 1 as an outcome. The probability curve obtained by the binomial logistic regression is sigmoid shaped, and in the range of 0 and 1(*Introduction to generalized linear models*, n.d.; Mathworks, n.d.; Seufert, 2014). The research conducted on AD where logistic regression have been applied would include choosing significant markers for the prediction of transformation from MCI to AD; also, the use of genetic algorithm with logistic regression for predicting transformations from healthy control to MCI/AD, and from MCI to AD (Johnson et al., 2014; Teipel et al., 2015).

## 1.7 Hypothesis

It is hypothesized that gene expression profiles of individuals with AD and healthy individuals would be different. Hence, even though there would be some genes having similar values of RNA levels for all samples; genes having distinctive RNA intensities in young samples, and in AD patients would be detected. Among those genes related with AD samples, novel biomarkers might be discovered for Alzheimer's disease. Besides, some of those genes might be involved in unknown pathways which might be related to the development of AD. In addition, there would be a correlation between RNA intensities of particular genes for young, old, and AD samples, and there might be a correlation between AD patients based on their RNA intensities. It might be possible to classify individuals according to their RNA levels, and to observe different clusters. Lastly, it would be possible to predict if an individual is healthy or diseased by identifying the genes related/not related with Alzheimer’s disease, and then using these genes for the prediction of AD risk.

# 2 METHODOLOGY & RESULTS

## 2.1 Methodology

### 2.1.1 Data collection and preprocessing

The dataset related with Alzheimer’s disease was downloaded from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/): GSE104704 which is an RNA-Seq dataset composed of total RNA isolated from frozen brain tissue samples (lateral temporal lobe) from 8 young healthy brains (Young), 10 aged healthy brains (Old), and 12 aged diseased brains (AD). The data were generated by Illumina NextSeq 500 (Homo sapiens).

Before the analysis, data with missing entries (rows containing values equal to zero) were removed. Then, the normalized data were obtained. The following analysis of the GSE104704 dataset was done by using Matlab.

### 2.1.1 Calculation of mean values and standard deviations

The first 200 genes of normalized data (after cleanup process, 128 genes were left) were used to calculate mean values of RNA intensities of each gene for all samples, for only young, for only old, and for only AD samples, respectively. These values were then used to find the genes having normalized RNA intensities as different for young and similar for old and AD samples, and also as different for AD and similar for young and old samples. To achieve these purposes, mean values of old samples were subtracted from mean values of young samples, mean values of AD samples were subtracted from mean values of young samples, and finally mean values of AD samples were subtracted from mean values of old samples. Maximum and minimum differences were found. Also, to find the genes having similar normalized RNA intensities for all samples, standard deviations of mean values were calculated. The minimum value was found. All of these values were taken into consideration for identifying genes with similar normalized RNA intensities, and distinct normalized RNA intensities. By using the mean values, bar plots of genes having similar normalized RNA intensities for all samples, genes having distinct normalized RNA intensities for young samples, and also for AD samples were generated. Two-sample t-test (young vs. old, young vs. AD, and old vs. AD, respectively) was performed to test if the mean values were equal or not. P-values (*p*) were also obtained to test the significance. In addition, standard deviations of young, old, and AD samples were calculated for each gene. To represent the variability of the data, standard deviation values were used to create error bars for each plot, and significance levels were indicated on all graphs.

### 2.1.2 Calculation of correlation coefficients and P-values

Correlation coefficients and P-values of genes were calculated to observe if there was a correlation between genes. This calculation was done for all, for only young, for only old, and for only AD samples, respectively. Correlation matrix of genes was created. Correlation coefficients greater than 0.9, and correlation coefficients lower than -0.9 were found for each calculation. Genes corresponding to these values were detected. Correlation graphs of correlated gene pairs among all samples were plotted. Also, to detect the correlation between samples, correlation coefficients and P-values of samples were calculated for only young, for only old, and for only AD samples, respectively. Correlation between samples was plotted. Correlation coefficient values greater than 0.9, and values smaller than -0.9 and their corresponding samples were found. A correlation graph of AD patients having correlated normalized RNA intensities was plotted. For each graph, linear regression analysis was done to form a regression line, and to find Pearson's correlation coefficient r with P-value. These were indicated on all graphs. The analysis above was performed by the first 200 genes of the GSE104704 dataset (after data cleanup, there were 128 genes left).

### 2.1.3 Principal component analysis and clustering

Principal component analysis (PCA) was performed to reduce the variable space and to represent the variance of the data with uncorrelated principal components, and to observe possible classifications. 1000 genes starting from the first row of the GSE104704 dataset were selected. After cleanup process, there were 661 genes left. The data were normalized. Two-sample t-test was performed between healthy (young and old) and diseased (AD) samples. P-values were obtained. These P-values were then added to the normalized data as the last column. The table containing gene IDs and the normalized data was sorted according to the P-values. Genes having P-values lower than or equal to 0.05 were selected (80 genes) and the used for the analysis. PCA was applied to all samples, for young and AD samples, and for old and AD samples, respectively. For each analysis, the matrix of principal components of the data, principal component scores, and the principal component variances are found. The number of principal components used for plotting the data was decided by the percentage of the cumulative sum of the principal component variances (>50%). For all principal component scatter plots, two principal component scores were used. In addition to score plots, time course of each principal component, and principal component variances were plotted. After PCA, cluster analysis of the same data was performed to group the individuals into clusters. By using the principal component scores, principal component plots with colored clusters were generated with the maximum number of clusters as 2 for all samples, young and AD samples, and old and AD samples, respectively. Finally, to visualize the hierarchical binary cluster tree of the normalized data, dendrogram plots were generated by using the optimal leaf order again for the all, young and AD, and old and AD samples.

### 2.1.4 Fitting data with binomial logistic regression model

Binary logistic regression was performed to estimate the probability of a patient having Alzheimer's disease based on that patient's normalized RNA intensities. For this purpose, same data used for PCA and clustering were also used (80 selected genes). The following procedure was applied to the first seven genes ($p$ < 5e-05), first three genes ($p$ < 3e-05), and last nine genes ($p$ between 0.004 and 0.005), respectively. Firstly, mean values of normalized RNA intensities of genes were calculated for each individual. Then, individuals were numbered from 1 to 30 (healthy= 1-18, diseased= 19-30), and sorted according to their average normalized RNA intensities. Certain intervals were determined for testing AD. Furthermore, a predictor matrix was created as the first column containing the number of AD patients and the second column containing the number of tested individuals in each interval. For the logistic regression, average normalized RNA intensities, predictor matrix, binomial distribution, and logit link were used. Logit coefficients with statistics were obtained, by using these values generalized linear model values were found. Proportion was calculated by dividing number of AD patients by number of tested individuals. Binomial logistic regression model was plotted by using average normalized RNA intensities, proportions, and generalized linear model values. Finally, a prediction code is created by using logit coefficients to display probability when an RNA level of a patient is input by the user.

### 2.1.5 Statistical tests

Statistical analysis of the data was performed by using Matlab. Two-sample t-test was used to test the significance of difference between means. Except where otherwise noted, P-values lower than 0.05 were considered to be significant and indicated by *** for values lower than 0.005, ** for values lower than 0.01 and, * for values lower than 0.05, respectively. P-values higher than 0.05 were considered to be non-significant and shown as N.S. In addition to P-values, for correlation graphs, Pearson's correlation coefficient r was calculated to show the strength of the relationship between variables (r close to ±1 indicates a strong linear relationship).

### 2.2 Results

Calculating and comparing the average values of normalized RNA intensities of 128 genes for young, old, and AD samples; three genes were found to have very similar normalized RNA levels in all samples. The first gene was KIAA0922 (or TMEM131L) gene having the mean normalized RNA intensities equal to 0.5196 for young, 0.5083 for old, and 0.5332 for AD samples. Differences were calculated as 0.0113 (*p* = .848) between young and old samples, as 0.0136 (*p* = .866) between young and AD samples, and as 0.0250 (*p* = .717) between old and AD samples (Figure 2.1A). The second gene was THRA gene whose average normalized RNA levels were 0.6814 for young, 0.6760 for old, and 0.6448 for AD samples. Differences between mean values were found as 0.0054 (*p* = .928) between young and old, as 0.0367 (*p* = .557) between young and AD, and as 0.0313 (*p* = .535) between old and AD samples (Figure 2.1B). The third gene was IDH1 gene having average normalized RNA intensity values for young, old, and AD as 0.5720, 0.5746, and 0.6052, respectively. Differences between average values of young and old, between young and AD, and between old and AD were calculated as 0.0026 (*p* = .972), 0.0332 (*p* = .705), and 0.0306 (*p* = .688) accordingly (Figure 2.1C).

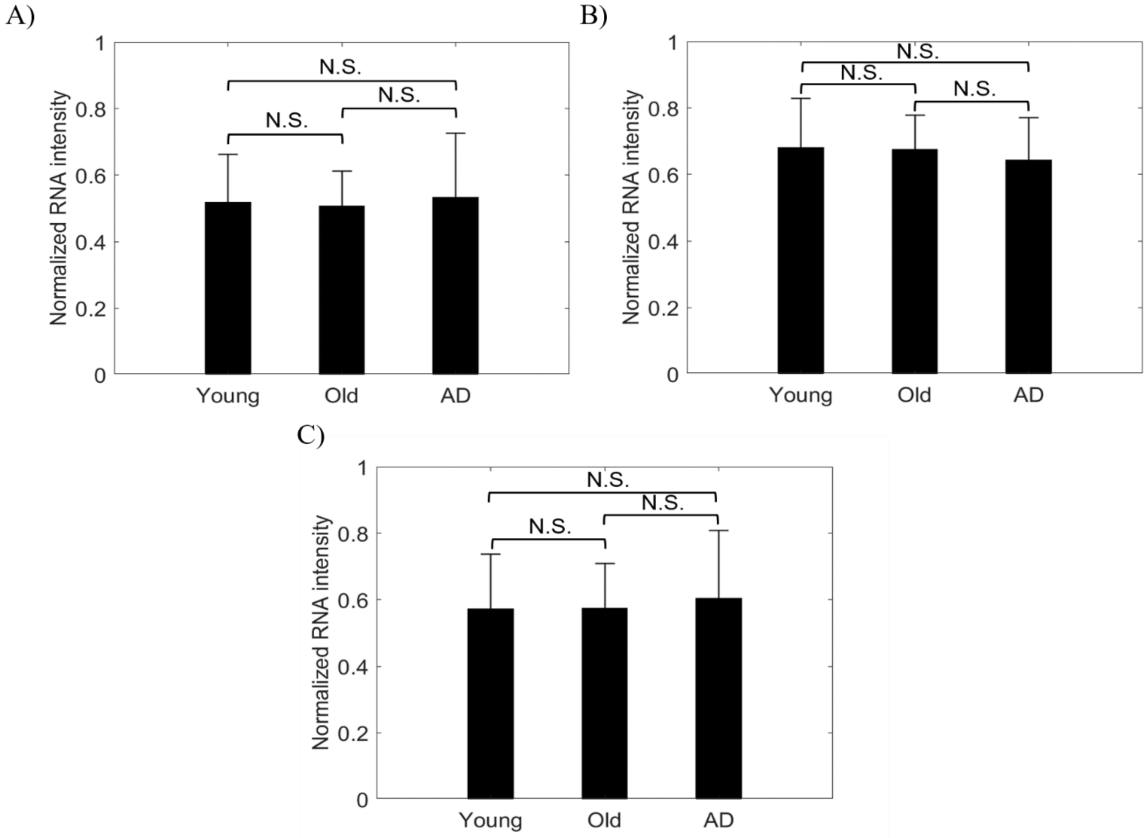

**Figure 2.1** Genes having similar normalized RNA intensities (average) in all sample groups. (A) KIAA0922 (or TMEM131L) gene. (B) THRA gene. (C) IDH1 gene.

Furthermore, two genes were found to be having distinct normalized RNA levels in young individuals. The first one was MECR gene having average normalized RNA levels as 0.4866 for young, 0.6443 for old, and 0.6451 for AD samples. The differences between RNA levels of young and old, and young and AD were calculated as 0.1577 ($p < .005$) and 0.1585 ($p = .016$), respectively. However, the difference between average RNA intensities of old and AD samples was very small: 0.0008 ($p = .988$). It might be observed that the RNA levels of old and AD samples were very similar, while there was a significant decrease in RNA levels of young samples (Figure 2.2A). The second one was PTPRD-AS2 gene. Average normalized RNA intensities of this gene for young, old, and AD were 0.2530, 0.4804, and 0.4837, respectively. Again, it was seen that the difference between RNA levels of old and AD was very small which was equal to 0.0033 ($p = .976$), while differences between young and old (0.2274, $p = .006$), and young and AD (0.2307, $p = .072$) were greater (Figure 2.2B). Even though the P-value for the difference between young and AD samples was found as greater than 0.05, it was accepted as significant.

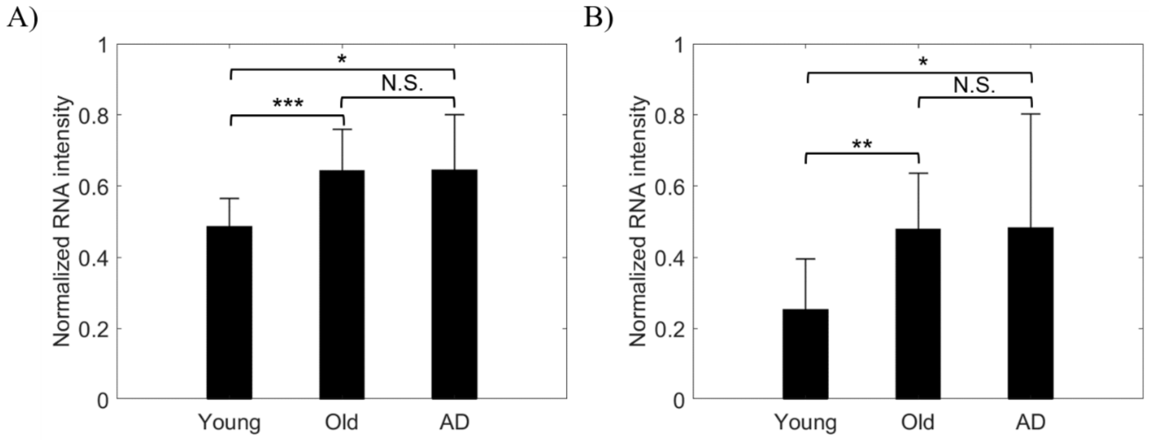

**Figure 2.2** Genes having different normalized RNA intensities (average) in young individuals. (A) MECR gene. (B) PTPRD-AS2 gene.

Moreover, three other genes were detected to have distinct normalized RNA intensities in individuals with AD. The first one was LOC283440 gene whose average normalized RNA level was 0.2570 in young individuals, 0.2946 in old indivudals, and 0.3581 in AD patients. The differences between young and AD and old and AD samples were calculated as 0.1281 ($p < .005$) and 0.0905 ($p = .044$), accordingly. However, the difference between young and old samples was calculated as 0.0377 ($p = .116$) which was really small compared to differences between other groups. The significant decrease in RNA levels of AD patients might be seen in the figure below (Figure 2.3A). The second gene was PRKACB that has average normalized RNA levels of 0.4327 for young, 0.4133 for old, and 0.7405 for AD samples. A significant increase was observed in normalized RNA levels of AD patients. It was detected that the normalized RNA levels of young and old individuals slightly differ from each other (0.0194, $p = .749$). Despite that, the differences between normalized RNA intensities of young and AD (0.3078, $p < .005$) and old and AD samples (0.3272, $p < .005$) were found to be highly significant (Figure 2.3B). The third gene was LINC01372 gene with the average normalized RNA intensity values of 0.8042, 0.7241, and 0.4680 for young, old, and AD samples, respectively. For this gene, a decrease in normalized RNA levels of AD patients was observed. The differences between average RNA levels of young and AD and old and AD samples were calculated as 0.3361 ($p < .005$) and 0.2561 ($p = .015$), correspondingly. However, the difference between mean values of normalized RNA intensities in young and old samples was calculated as 0.0801 ($p = .365$) which was a smaller value as compared to the differences between other sample groups (Figure 2.3C).

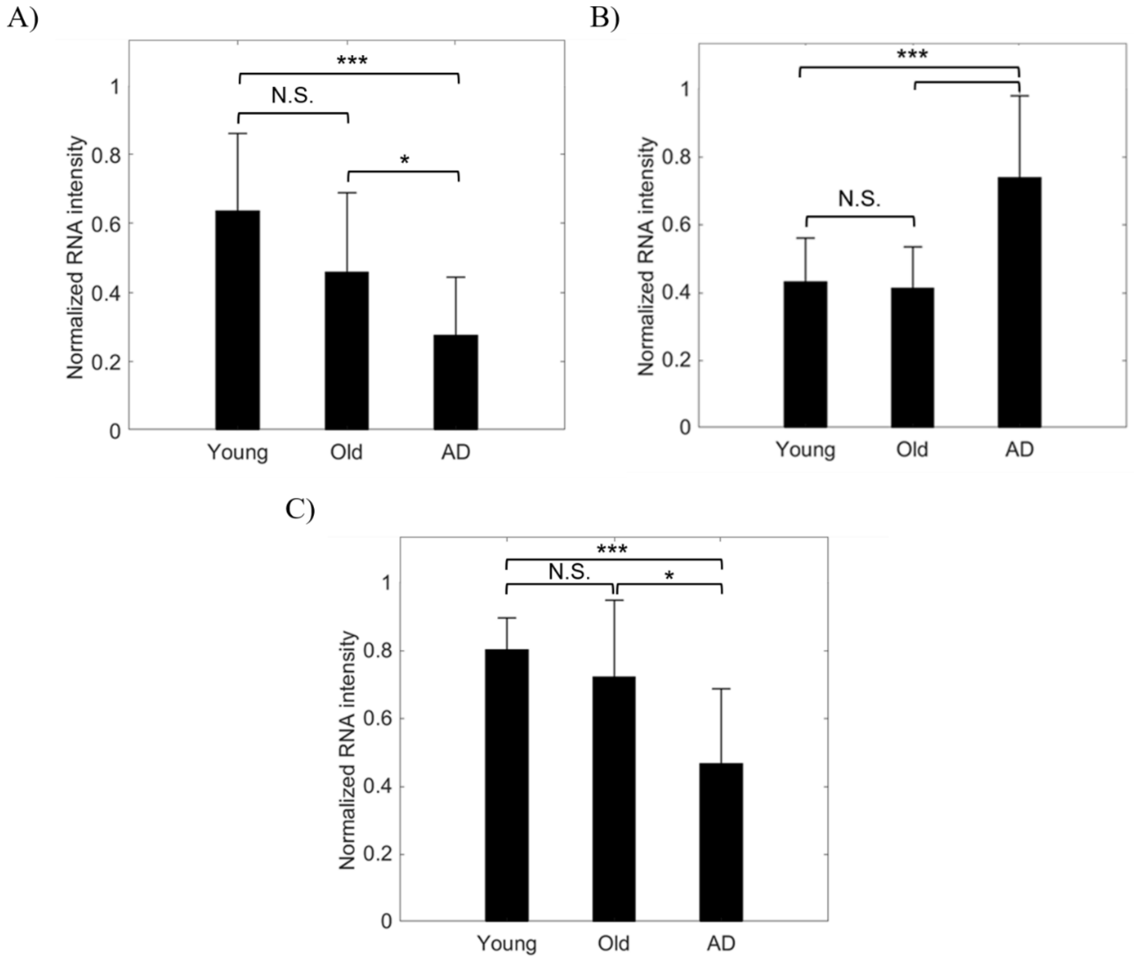

**Figure 2.3** Genes having different normalized RNA intensities (average) in AD patients. (A) LOC283440 gene. (B) PRKACB gene. (C) LINC01372 gene.

By the calculation of correlation coefficients (r) and P-values of 128 genes for all samples, a correlation matrix was created, and four gene pairs were detected to have highly correlated normalized RNA intensities in all individuals. In correlation matrix, strongly correlated genes were demonstrated as dark red, or dark blue whose correlation coefficient values were near +1 or -1, respectively. Gene pairs having correlation coefficient values between 0.9 and -0.9 were regarded as uncorrelated or poorly correlated, and shown with other colors (shades of green, turquoise, yellow, orange, light red, and light blue). Even though several data points colored in red and blue might be observed, the majority of the correlation matrix was colored in green, turquoise, and yellow which indicates a poor correlation. The correlation matrix can be seen in the figure below (Figure 2.4).

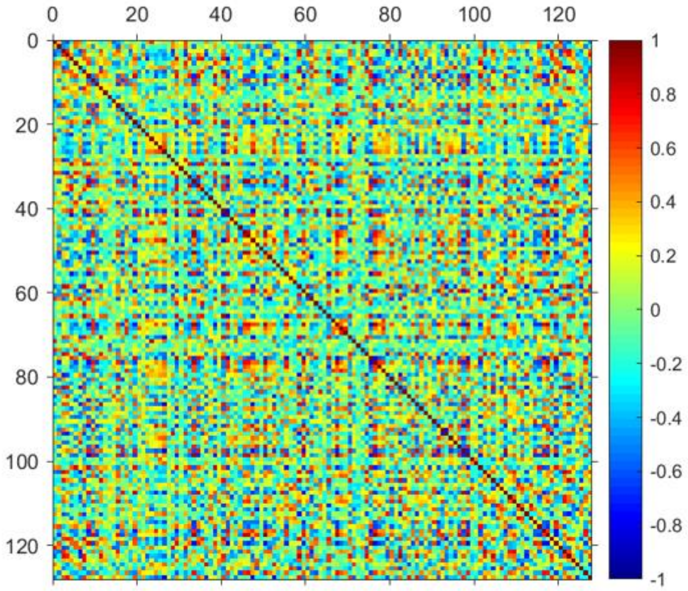

**Figure 2.4** Correlation between normalized RNA intensities of 128 genes in all individuals. Correlation matrix was created by using correlation coefficients. x-axis and y-axis contain genes starting from 1 to 128. Colors corresponding to correlation coefficient values can be seen in the colorbar at the right side. Values near ±1 indicate a strong linear relationship, whereas values equal to and near zero indicate that there is no correlation, or a very poor correlation.

Among the correlated genes, four gene pairs which had the highest r values (higher than 0.9) were detected and plotted in decreasing order of r values as the first one having the highest r value, whereas the latest one having the smallest r value. Relationships between normalized RNA intensities of pairs of genes can be seen in the figure below (Figure 2.5). The higher the r value, the stronger the linear relationship between genes. The first gene pair which can be seen in Figure 2.5A was SFR1 and CPPED1 (118th and 51st genes, respectively) with an r value of 0.9713 ($p$ < .005). The second gene pair was CPPED1 and SGK1 (51st and 34th genes, respectively) with an r value of 0.9465 ($p$ < .005) and can be seen in Figure 2.5B. The third gene pair was DOCK5 and ATP10B (68th and 77th genes, respectively) with an r value of 0.9366 ($p$ < .005) and can be observed in Figure 2.5C. The last gene pair was SFR1 and SGK1 (118th and 34th genes, respectively) with an r value of 0.9250 ($p$ < .005) and can be seen in Figure 2.5D. Considering the fact that all P-values were lower than 0 .005, it might be said that those genes significantly correlate with each other in young, old, and AD samples.

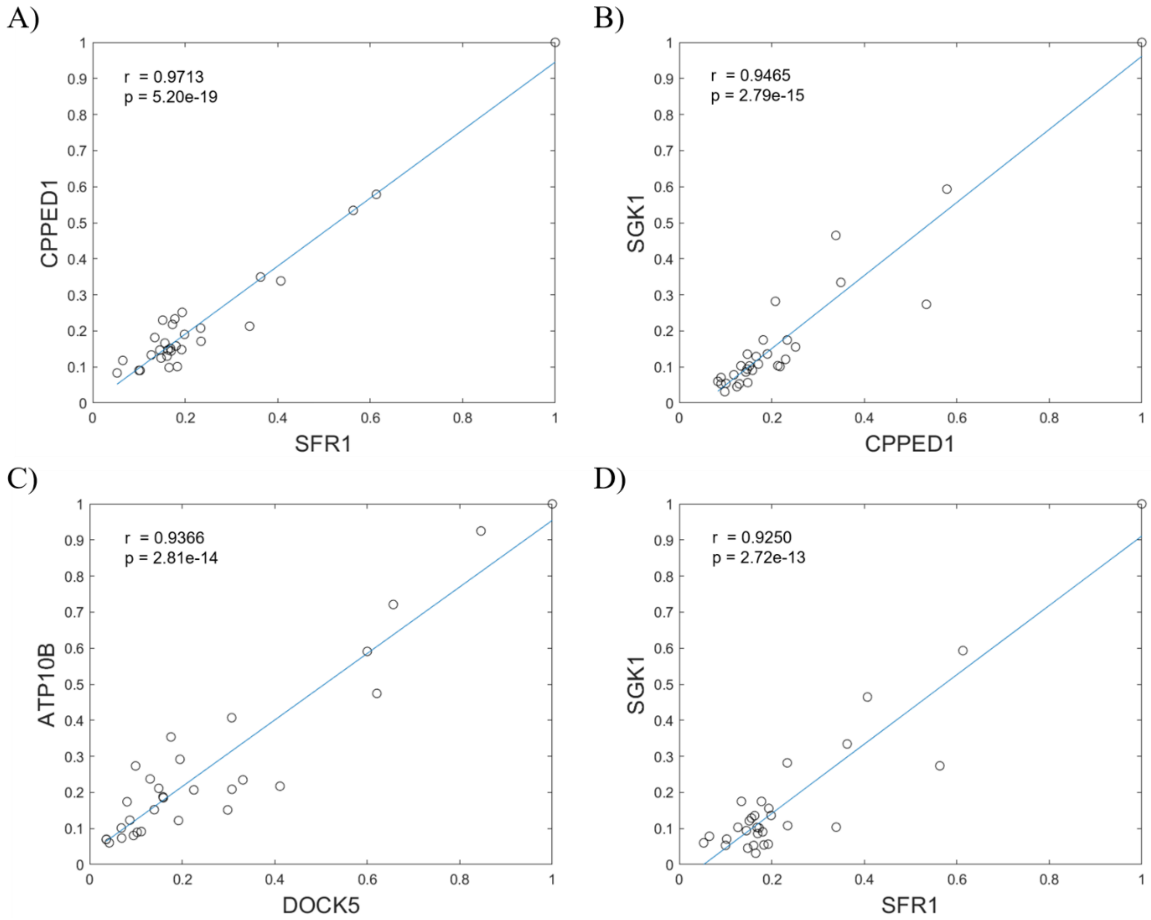

**Figure 2.5** Correlation between normalized RNA intensities of gene pairs. Correlation coefficients were calculated for all samples (young, old and AD). Top 4 gene pairs that had the highest Pearson correlation coefficients (r) were plotted. P-values and linear regression lines were also shown. (A) Correlation between SFR and CPPED1 genes. (B) Correlation between CPPED1 and SGK1 genes. (C) Correlation between DOCK5 and ATP10B genes. (D) Correlation between SFR1 and SGK1 genes.

Moreover, by the calculation of correlation coefficients and P-values of samples for only young, only old, and only AD, it was found that there was a strong linear relationship between the 3rd and the 7th AD patients based on their normalized RNA intensities. The correlation coefficient for these patients was calculated as 0.9092 ($p < .005$) which was really high and indicating a strong correlation, whereas the maximum r value found for young individuals was 0.7402, and for old individuals was 0.7399 which were not as high as the r value of 3rd and 7th patients. So, only the correlation graph of these two patients was plotted (Figure 2.6).

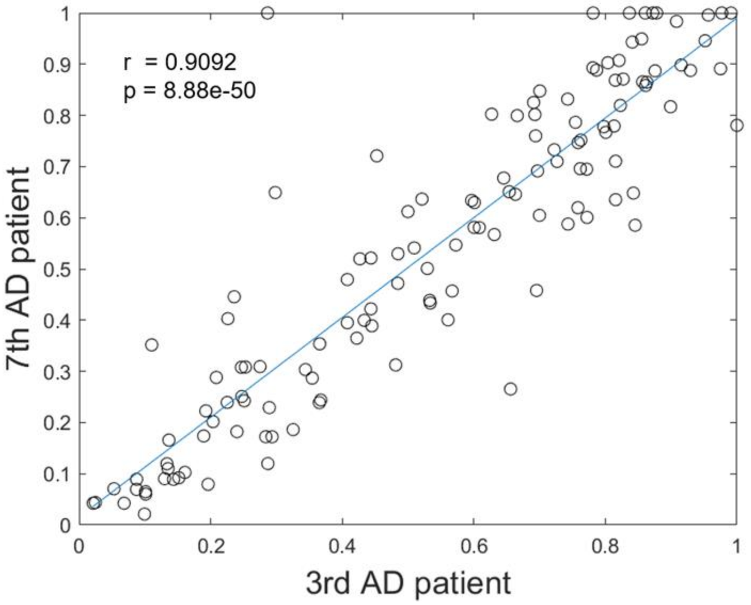

**Figure 2.6** Correlation between 3$^{rd}$ and 7$^{th}$ AD patients' normalized RNA intensities. Each gene was taken into consideration for calculation of Pearson correlation coefficients (r). r value, P-value and the linear regression line were indicated on the graph.

Among the first 1000 genes of the GSE104704 dataset, 80 genes were selected according to the P-values (*p* values < .05) obtained by two-sample t-test. The selected genes are shown in the table below (Table 2.1). PRKACB, LINC01372, and LOC283440 genes (highlighted in orange in Table 2.1) which were found to have distinct RNA levels in AD samples were also detected in the selected genes. The first seven genes, and the last nine genes that were used for binary logistic regression analysis were highlighted in yellow and green, respectively (Table 2.1).

**Table 2.1** The list of 80 selected genes detected by performing two-sample t-test between healthy (young and old) and AD samples' normalized RNA intensities.

| Gene ID | Gene ID | Gene ID | Gene ID | Gene ID |
|---|---|---|---|---|
| 1- LAPTM4B | 17- NDUFB6 | 33- DSCAM-AS1 | 49- C9orf72 | 65- RGL3 |
| 2- CXorf56 | 18- EFL1 | 34- NFU1 | 50- EIF3J | 66- FOXRED1 |
| 3- RTCA | 19- AASDHPPT | 35- GBGT1 | 51- NOTUM | 67- G2E3 |
| 4-TSPAN5 | 20- KANK2 | 36- BCL9L | 52- HIC2 | 68- GPR137C |
| 5-UTP11 | 21- PLIN4 | 37- ZRSR2 | 53- JKAMP | 69- GPALPP1 |
| 6- ATXN10 | 22- KLK14 | 38- SNRPD2 | 54- ELP6 | 70- COLCA2 |
| 7- PRKACB | 23- VPS35 | 39- GRPEL1 | 55- LOC283440 | 71- LOC100507002 |
| 8- C11orf58 | 24- LINC01372 | 40- AAGAB | 56- PRR25 | 72- CIAO1 |
| 9- MICALL2 | 25- TUBA3D | 41- CMSS1 | 57- SERPIND1 | 73- TCEA3 |
| 10- LAMTOR5 | 26- COL28A1 | 42- NWD1 | 58- CWF19L1 | 74- LIAS |
| 11- C5orf30 | 27- HDAC10 | 43- MYO7A | 59- EFCAB12 | 75- HCFC2 |
| 12- UROD | 28- CAPZB | 44- ATP5C1 | 60- ADH5 | 76- DLG2 |
| 13- NUDT21 | 29- CRHR1-IT1 | 45- CCNDBP1 | 61- B4GALT1 | 77- TRMT10C |
| 14- EIF2S1 | 30- TMX2 | 46- REPS2 | 62- LINC00951 | 78- PTS |
| 15- ATP6AP2 | 31- PCYT1B | 47- ZFP36L1 | 63- SNRPB | 79- LOC100133077 |
| 16- AREL1 | 32- ZRANB3 | 48- COPA | 64- TKFC | 80- PLAC4 |

PCA was applied to the normalized RNA intensities of 80 genes selected among the first 1000 genes of the GSE104704 dataset (*p* values < .05) for observing possible classifications. Considering the analysis of all samples (young, old, and AD), it might be said that the classification was not very clear. Most of the AD patients were observed at the left side of the plot except for the 20th, 27th, and 30th individuals. The 20th individual was observed at the right side. Young and old samples were observed as intermixed, instead of forming separate distinct classes. Young individuals and old individuals were observed at the right side of the plot except for the 2nd, 4th, 13th, and 18th individuals. The 18th individual was seen at the left side. The 2nd and 4th individuals who were young, and 13th individual who was old, and 27th and 30th individuals who were AD patients were observed in the middle of the plot as intermixed (Figure 2.7A).

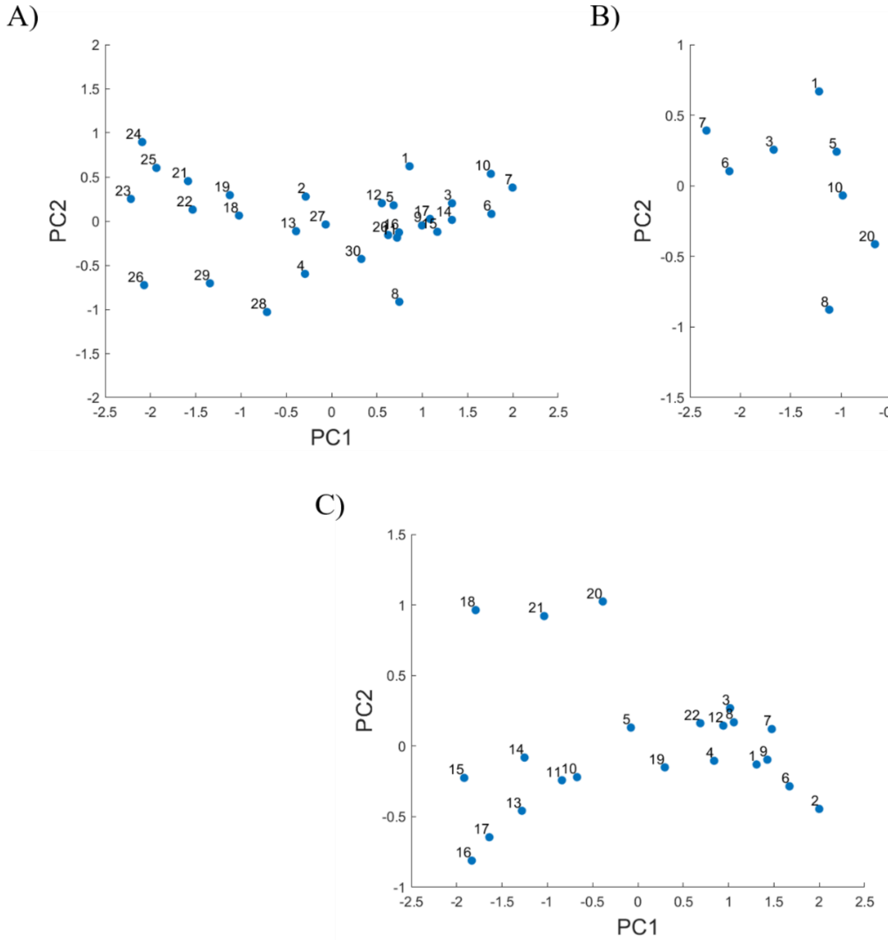

**Figure 2.7** Principal component analysis of normalized RNA intensities of individuals. The first and the second principal components were indicated as "PC1" and "PC2", respectively. Numbers near the data points correspond to the individuals. (A) PCA plot of all samples (Young: 1-8, Old: 9-18, and AD: 19-30). (B) PCA plot of young and AD samples (Young: 1-8, AD: 9-20). (C) PCA plot of old and AD samples (Old: 1-10, AD: 11-22).

Regarding the analysis of young and AD samples, again there was no obvious classification of samples. Young individuals were observed at the left side of the plot apart from the 2nd and 4th individuals, while AD patients were located at the right side of the plot excluding the 10th, 17th, and 20th individuals. The 10th individual was observed at the left side. Individuals numbered as 2, 4, 17, and 20 were located in the middle of the plot which consist of both young individuals and individuals with AD (Figure 2.7B). Individuals numbered as 10, 17, and 20 were the same individuals with the individuals numbered as 20, 27, and, 30 in Figure 2.7A, respectively. After the analysis of old and AD samples, it was seen that most of the old individuals were located at the right side of the plot except for the 5th and 10th individuals. The 5th individual was observed in the middle of the plot, while 10th individual was located at the left side. The majority of the AD patients were observed at the left side of the plot aside from the 12th, 19th and 22nd individuals which were located at the right side (Figure 2.7C). The 5th, 10th, 12th, 19th, and 22nd individuals were the same individuals with 13th, 18th, 20th, 27th, and 30th indivuals in Figure 2.7A, respectively. Regarding the data set containing 30 individuals in total, it might be said that the 2nd and 4th individuals who were young; 13th and 18th individuals who were old; and 20th, 27th and 30th individuals who were AD patients were detected at different sides of the plot than the sides where they would normally be located at.

Moreover, the time course of the first 12 principal components were obtained from PCA of all samples, young and AD, and old and AD samples (Figure 2.8). It was observed that the data was noisier for the first principal components, and the noisiness was becoming less and less from the first PC to the last PC in all plots PCA results of all groups shown that the PC scores for the first PCs were between -2 and 2; however, the interval for PC scores of the second PCs were between -1 and 1, and the PC scores of last principal components were around zero. The time course of first 12 PCs obtained from PCA of all samples (Figure 2.8A), PCA of young and AD samples (Figure 2.8B), and PCA of old and AD samples (Figure 2.8C) can be seen below.

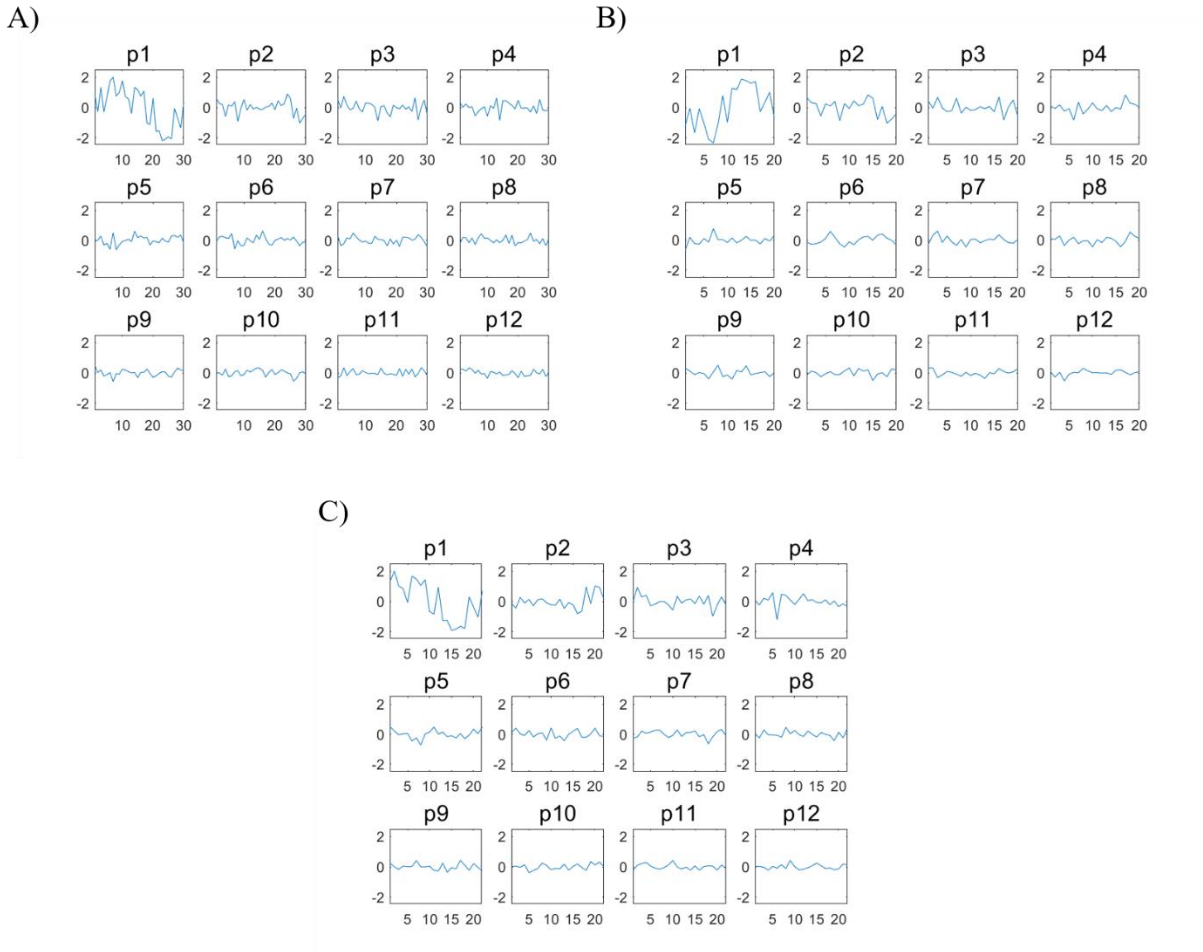

**Figure 2.8** The time course of first 12 principal components. x-axis contains the PC numbers, while y-axis contains the PC scores. (A) PCs obtained from PCA of all samples. (B) PCs obtained from PCA of young and AD samples. (C) PCs obtained from PCA of old and AD samples.

As a result of PCA analysis, variance plots were also obtained (Figure 2.9). According to these plots, it might be said that the first principal components were holding the highest variances, and for each analysis the most part of the data variability was explained by the first two PCs. After these PCs, there was a sharp decrease in the variance.  For the analysis of all samples, the variance of the first PC was found as 1.6548 (59.60%), and the variance of the second PC was found as 0.2114 (7.61%) which together explained the 67.21% of the total variance (Figure 2.9A). For the analysis of young and AD samples, the first PC’s variance was found as 1.8372 (61.02%), while the second PC’s variance was found as 0.3055 (10.15%). These two PCs explained 71.17% of the total variance (Figure 2.9B). Lastly, for the analysis of old and AD samples, the variance of the first PC was found as 1.6934 (61.96%), and the variance of the second PC was found as 0.2282 (8.35%). Totally, their contribution to the total variance was 70.31% (Figure 2.9C). Therefore, the maximum value of variance was observed as higher for both the first and second principal components obtained by PCA of young and AD samples.

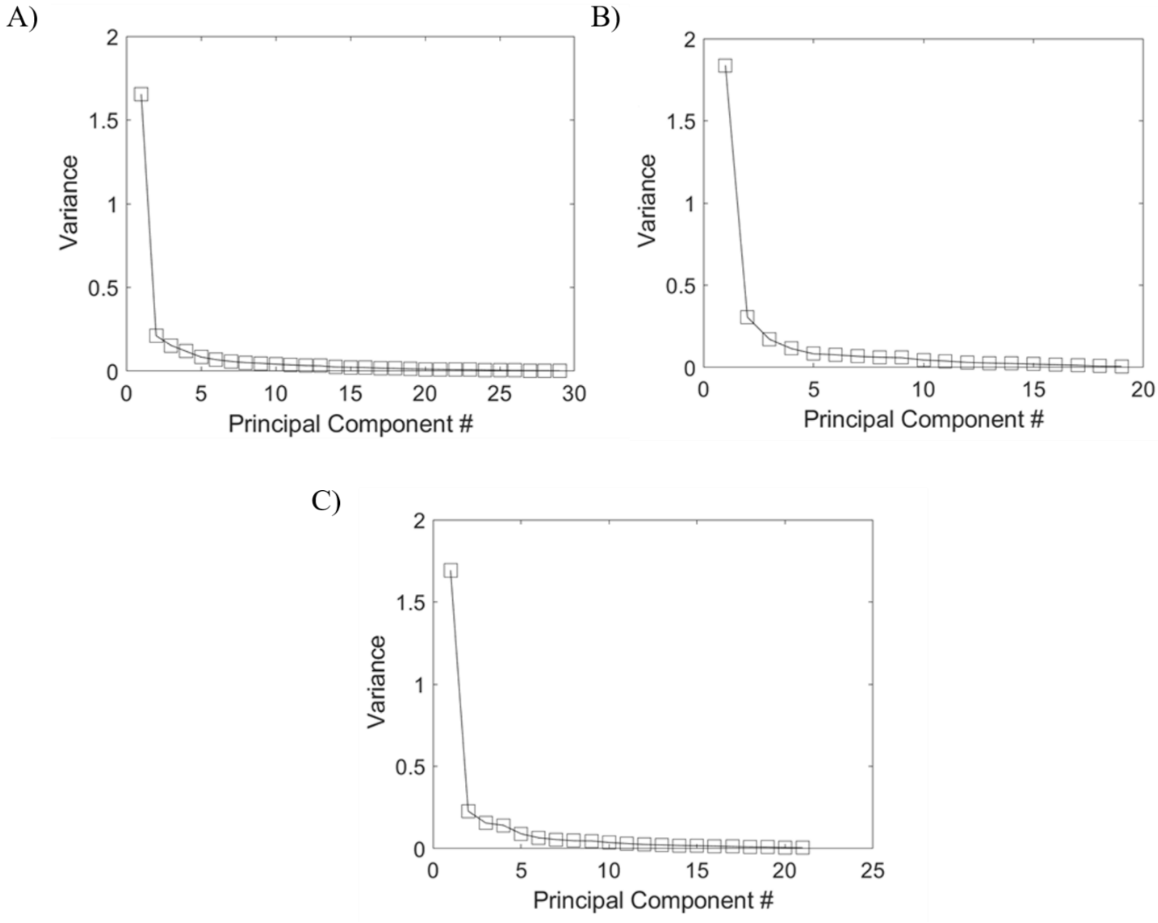

**Figure 2.9** Variances of each principal component. (A) PCs obtained from PCA of all samples. (B) PCs obtained from PCA of young and AD samples. (C) PCs obtained from PCA of old and AD samples.

By the clustering analysis of normalized RNA intensities of 80 genes selected among the first 1000 genes of the GSE104704 dataset (*p* values < .05) for all samples, young and AD, and old and AD samples, two clusters were observed for each analysis. For the analysis of all samples, it was observed that individuals with AD except for the 20th, 27th, 28th, and 30th ones formed a cluster at the left side of the plot (points shown in red). However, among those individuals there was also the 18th individual who was an old and healthy individual. On the other hand, young and old individuals except for the 18th indiviudal formed another cluster at the right side of the plot (points shown in blue). The 20th, 27th, 28th, and 30th individuals who were AD patients were also observed in this cluster (Figure 2.10A). For the analysis of young and AD samples, it was seen that all young individuals formed a cluster at the left side of the plot (points shown in red), however this cluster was containing also individuals with AD (10th, 17th, and 20th individuals). AD patients other than those formed the second cluster at the right side and they were shown in red (Figure 2.10B). These individuals were the same individuals who were numbered as 20, 27, and 30 in Figure 2.10A.

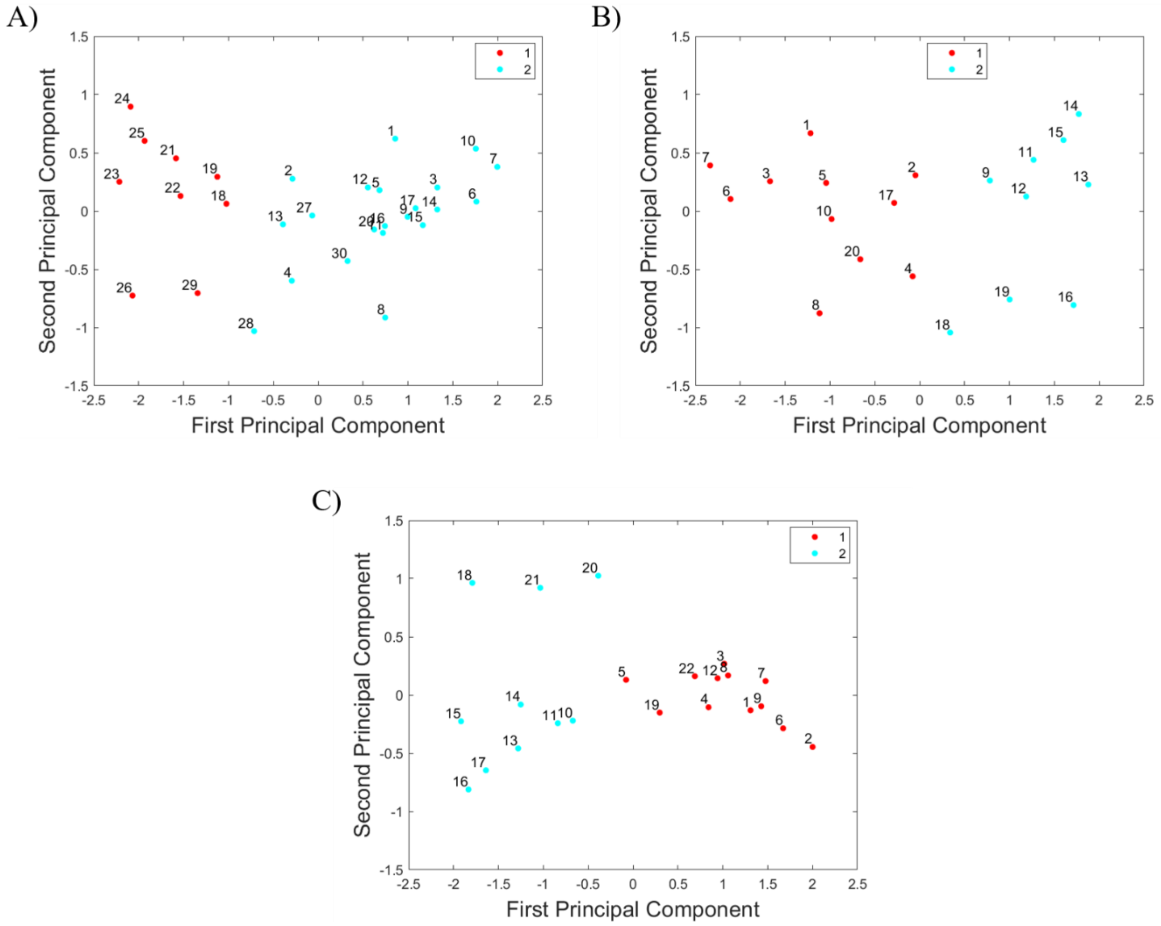

**Figure 2.10** Cluster analysis of normalized RNA intensities of individuals. Two clusters were formed. Numbers near the data points correspond to the individuals. (A) Cluster plot of all samples (Young: 1-8, Old: 9-18, and AD: 19-30). (B) Cluster plot of young and AD samples (Young: 1-8, AD: 9-20). (C) Cluster plot of old and AD samples (Old: 1-10, AD: 11-22).

For the analysis of the old and AD samples, it was noted that old indiviudals except for the 10th individual formed a cluster at the right side of the plot (points shown in red). Although, AD patients numbered as 12, 19 and 22 were also contained in this cluster. These individuals were the same individuals who were denoted as 20, 27, and 30 in Figure 2.10A. Other individuals with AD clustered at the left side of the plot (points shown in blue), and the 10th individual who was old were also observed in this cluster (Figure 2.10C). The 10th individual was the same individual who was numbered as 18 in Figure 2.10A.

Hierarchical cluster analysis was applied to the normalized RNA levels of 80 genes selected among the first 1000 genes of the GSE104704 dataset (*p* values < .05) for all samples, young and AD, and old and AD samples. Dendrogram plots containing two main clusters were obtained (Figure 2.11). For the dendrogram plot of all samples, it was noted that except for the 20th, 27th, and 30th individuals, AD patients and their corresponding leaf nodes were located at the right side of the dendrogram, whereas all of the young and old individuals and their related leaf nodes were located at the left side of the dendrogram including the AD patients who were numbered as 20, 27, and 30. The distance between two clusters was found as 2.80 (Figure 2.11A).

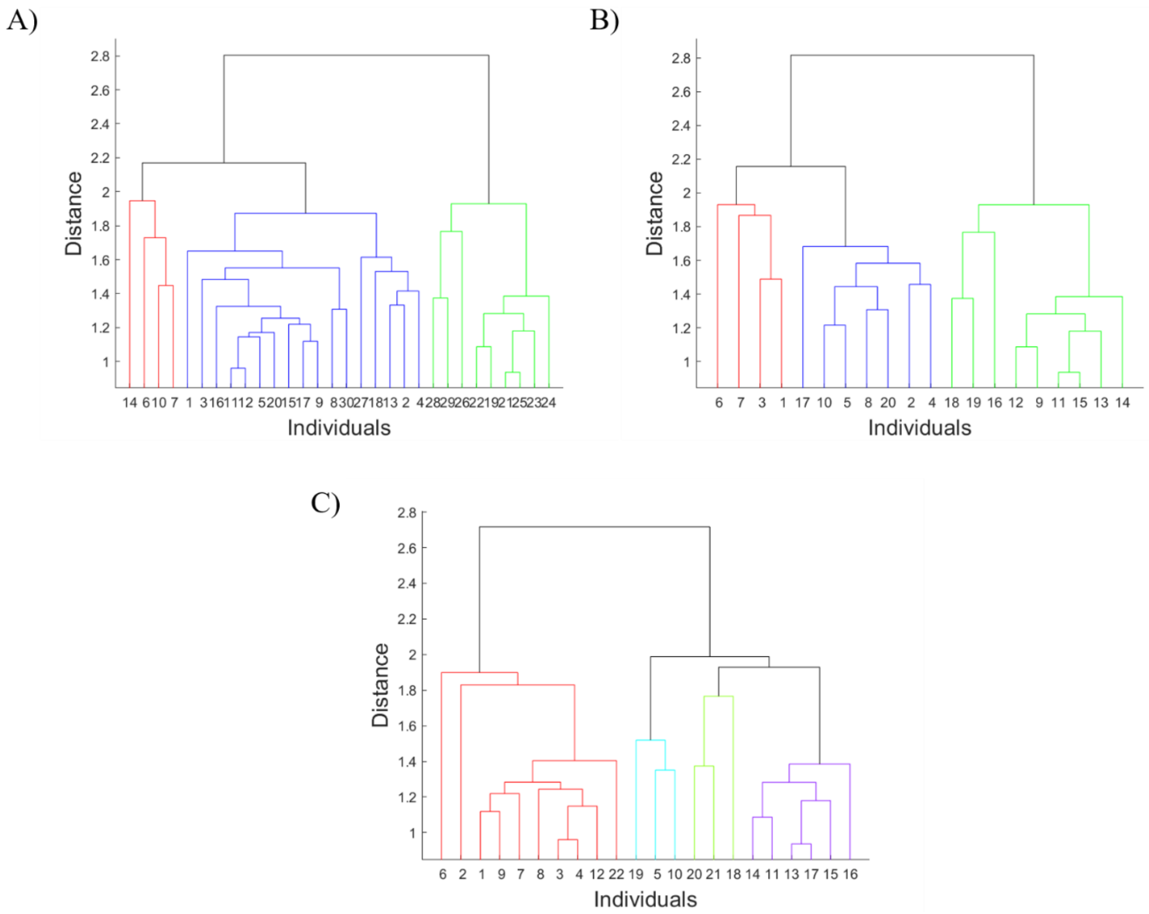

**Figure 2.11** Dendrogram plots of individuals obtained from hierarchial cluster analysis of normalized RNA intensity data. x-axis represents the individuals and y-axis correspond to the distance between clusters. The smaller the distance, the higher the similarity between individuals. (A) Dendrogram plot of all individuals (Young: 1-8, Old: 9-18, and AD: 19-30). (B) Dendrogram plot of young and AD samples (Young: 1-8, AD: 9-20). (C) Dendrogram plot of old and AD samples (Old: 1-10, AD: 11-22).

For the dendrogram plot of young and AD samples, it was observed that the first cluster was formed at the left side of the dendrogram by young individuals. However, some AD patients (10th, 17th, and 20th individuals) were also detected in this cluster. These individuals were the same individuls who were denoted as 20, 27, and 30 in Figure 2.11A. The rest of the AD patients formed the second cluster at the right side of the dendrogram. The distance between these clusters was calculated as 2.81 (Figure 2.11B). Finally, the dendrogram plot of old and AD samples shown that old individuals except the 5th individual clustered at the left side of the dendrogram, whereas AD patients except the 12th and 22th individuals clustered at the right side of the plot. The 5th, 12th, and 22th individuals were the same individuals with the 13th, 20th, and 30th individuals in Figure 2.11A. The distance between two main clusters were found as 2.71 (Figure 2.11C).

Binary logistic regression analysis was performed to predict if a patient has the Alzheimer's disease based on the normalized RNA intensity values of that patient. Binary logistic regression models were obtained. The first model was obtained by analyzing the average normalized RNA intensities of the first seven genes. An S-shaped curve was observed. The average normalized RNA levels were in the range 0.4-1.0. Smaller average normalized RNA intensity values were observed at the bottom part of the model, while greater values were observed at the top. For this model, P-values for logit coefficients were found as 0.0020 and 0.0029, respectively (Figure 2.12A).

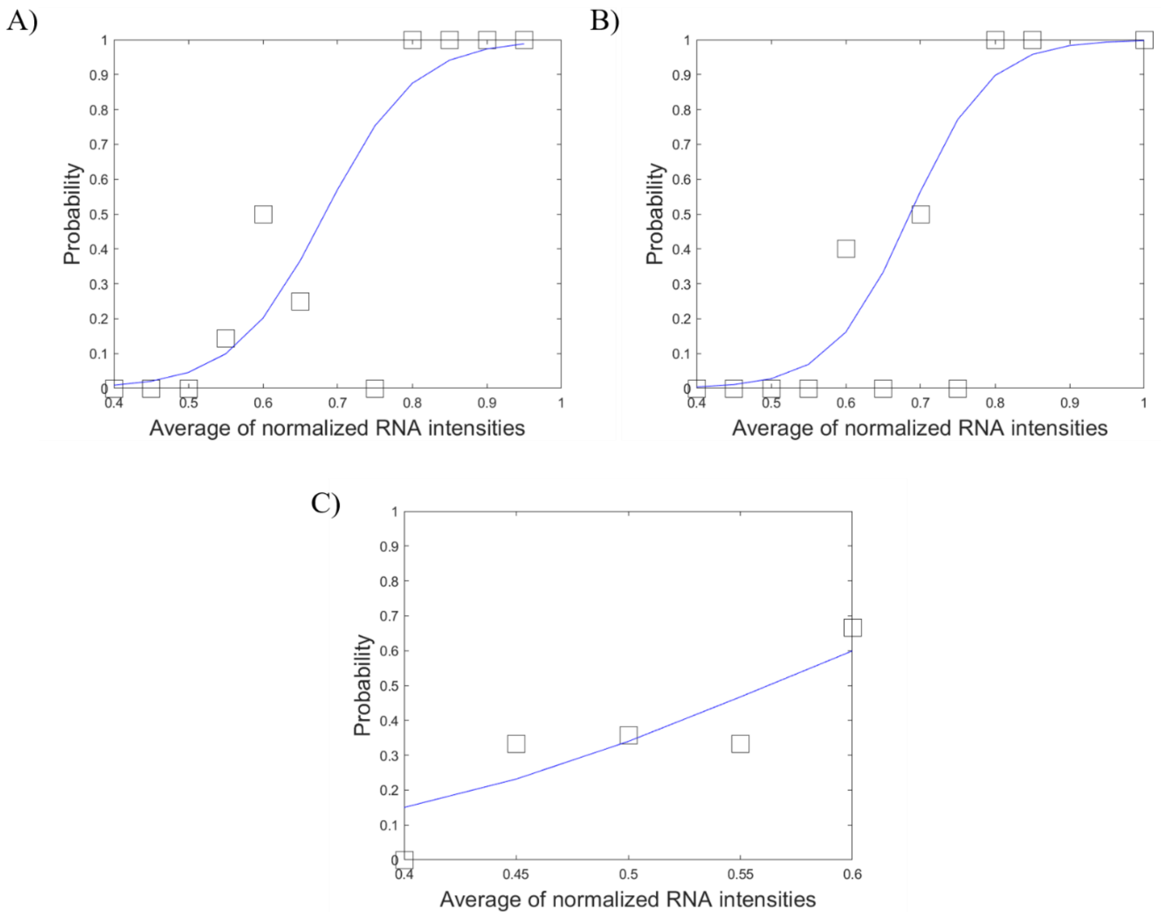

**Figure 2.12** Binary logistic regression models indicating the probability of a patient's having the Alzheimer's disease. (A) Model fit to the data which consists of the first seven genes ($p <$ 5e-05). (B) Model fit to the data which consists of the first three genes ($p <$ 3e-05). (C) Model fit to the data which consists of the last nine genes ($p$ between 0.004 and 0.005).

The second model was obtained by the analysis of the average normalized RNA levels of the first three genes. Again, an S-shaped curve was detected, and the average normalized RNA intensity values lied between 0.4 and 1.0. Smaller values of average normalized RNA intensity were located at the bottom part of the figure, whereas larger values were found at the top part. P-values for logit coefficients were calculated as 0.0021, and 0.0026, respectively (Figure 2.12B). The last model was obtained by fitting the data which was composed of the last nine genes. Instead of an S-shaped curve, a line was observed, and the average normalized RNA intensity values lied between 0.4 and 0.6. For this model, P-values for logit coefficients were found to be higher than 0.05 (Figure 2.12C).

The prediction code was created for each model to calculate and display the probability of AD risk. When an average normalized RNA intensity value of 0.9 was entered, the first model's code returned 97.4%, the second model's code returned 98.3%, and the last model's code returned 97%. When the value of 0.5 was entered, the results for the 1st, 2nd, and the 3rd models' codes were 4.5%, 2.8%, and 33.9%, respectively.

# 3 DISCUSSION & CONCLUSION

## 3.1 Discussion

The nature of Alzheimer's disease is highly complex, and several prognostic and predictive biomarkers have been detected to date. However, it is still required to discover novel biomarkers by identifying the genes associated with AD. In this study, RNA-seq data was used to detect the genes related with AD by comparing the average normalized RNA intensity levels of genes in young, old, and AD samples. It was expected to observe various genes having similar levels of RNA in all samples, and different levels of RNA in young, and in AD samples. As expected, KIAA0922 (or TMEM131L), THRA, and IDH1 genes were detected to have very similar RNA intensities in all samples, whereas RNA levels of MECR and PTPRD-AS2 genes differed significantly in young individuals, and LOC283440, PRKACB, and LINC01372 genes were observed to have very distinct RNA intensities in individuals with AD (See Figures 2.1, 2.2, 2.3). It was deduced that KIAA0922, THRA, and IDH1 genes had no association with AD, since their RNA levels were similar in all samples. THRA gene is involved in superpathways such as gene expression, nuclear receptor transcription, and IDH1 gene in carbon metabolism, citrate cycle, glucose metabolism, amino acid metabolism, etc. (Belinky et al., 2015). Additionally, KIAA0922 is a protein coding gene which is responsible for regulating the proliferation and differentiation within the thymus, and is one of the known antagonists of the Wnt signaling pathway (Maharzi et al., 2013). The fact that these pathways are not included in the pathways that are involved in the development of AD confirms the deduction (Giri et al., 2016). On the contrary, genes that have distinct RNA levels in young individuals and in AD patients might be involved in pathways that are associated with AD pathology. MECR gene have been found to be involved in fatty acid metabolism, and in their study, Snowden et al. (2017)  suggests that there is an important dysregulation of unsaturated fatty acid metabolism in AD brains  (Belinky et al., 2015). Also, in their research, Schipper, Maes, Chertkow, and Wang (2007) have denoted MECR as one of the predicted miRNA targets which have lower levels in AD. However, there is no pathway information available for PTPRD-AS2 gene. In addition, the most significant risk factor for developing AD is age. The age-associated changes that are occurring in old individuals contain atrophy, inflammatory response, vascular damage, production of free radicals, and failure in production of energy inside the cells. These changes promote the development of AD (National Institutes of Health, 2019b). Therefore, similar RNA levels might be observed in aged and AD brains. This would explain the similarity between RNA levels of MECR and PTPRD-AS2 genes in old and AD samples. Moreover, for the genes that had distinct RNA levels in AD patients it has been found that one of the superpathways that PRKACB gene is involved in is DAG and IP3 signaling which contains the neurophysiological process PGE2-induced processing of pain, and by great number of studies it is suggested that altered pain levels are observed in AD patients as a result of variations in pain processing (Belinky et al., 2015; Defrin et al., 2015). Therefore, the distinct RNA levels in AD patients for PRKACB gene might be explained by its involvement in that particular pathway. However, LOC283440 gene is an uncharacterized gene, and no pathway data was available for both LOC283440 and LINC01372 genes.

Besides, correlation coefficients and corresponding P-values were calculated for detecting the relationships between genes, and between AD patients. It was expected to observe several correlated gene pairs in all individuals. Also, RNA intensities of AD patients would be correlated, since alterations observed in gene expression of AD patients are similar (Bottero & Potashkin, 2019). As expected, numerous gene pairs were found to be positively correlated in all samples, and 4 pairs were reported (See Figures 2.4, 2.5). The correlation between genes might indicate that they are involved in the same pathway. However, there was no pathway data available to support this. Also, it was found that 3rd and 7th AD patients were showing a strong positive correlation based on their RNA levels, as it was expected (See Figure 2.6). This finding indicates that these patients' RNA intensities of each gene were very similar to each other when compared to other AD patients' RNA levels. Considering the results obtained by the comparison of RNA levels and correlation analysis, it might be said that if the analyses had been performed by using greater number of samples and much more genes instead of using only the first 200 genes, and by using the genes with higher significance levels for detecting the differences between healthy individuals and AD patients, better results might have been obtained. Therefore, before performing PCA, cluster analysis, and binary logistic regression analysis, two-sample t-test was applied between healthy and diseased samples' RNA levels of first 1000 genes. The selected genes were reported in Table 2.1. They were the most important genes for differentiating between healthy and AD samples. LOC283440, PRKACB, and LINC01372 genes were found to be included in these genes. This finding was consistent with the results found by previous analyses, since they were also detected as the genes having distinct RNA levels in AD patients. The first three genes of selected genes were LAPTM4B, CXorf56, and RTCA. They might be involved in pathways associated with AD pathogenesis. The LAPTM4B gene has been found to be involved in lysosome pathway (Belinky et al., 2015). Several studies state that lysosome dysfunctioning is significantly involved in AD development, and lysosomes are engaged in both amyloid beta formation and their toxic effects to the nerve cells (Kukar, 2019; Zheng, 2012). Moreover, in their study, Pollock et al. (2019), report that the protein product of CXorf56 gene is possibly involved in connecting APOE4 to activate microglia, and mediate the inflammatory response related to Alzheimer's disease. Although there was no pathway data available for RTCA gene, in his study, Izgi (2017) states RTCA gene among the downregulated AD-related genes. Therefore, the literature confirms the association of LAPTM4B, CXorf56, and RTCA genes with AD.

Principal component analysis was performed to visualize different classes formed by different sample types. It has been widely preferred method in numerous AD studies as mentioned in the introduction chapter. The PCA plot of young, old, and AD samples showed that there was no distinct classification between young and old samples; however, AD samples and young/old samples were observed at different sides of the plot (See Figure 2.7A). These findings indicate the similarity between RNA levels of healthy individuals (young and old); also, the altered RNA levels in AD patients. For the PCA plot of young and AD samples; the separation was not distinct, but young and old individuals were observed at different sides of the plot with some exceptions (See Figure 2.7B). Again, it indicates the difference between RNA intensities in young and old samples. Likewise, in the PCA plot of old and AD samples, old samples were detected at the right side, whereas AD patients were detected at the left side indicating the difference between their RNA levels (See Figure 2.7C). The following individuals were enumerated according to the data set (Young: 1-8, Old :9-18, AD: 19-30). The exceptions detected by the principal component analysis (2nd, 4th, 13th, 18th, 20th, 27th, and 30th individuals) might mean the misdiagnosis of these individuals. The 2nd and 4th individuals were denoted as healthy young individuals; however, in both PCA plots of all samples, and young and AD samples, they were observed in the middle of the plot intermixed with other samples. So, their RNA levels have shown similarity between RNA levels of AD samples. Also, despite the fact that the 13th and 18th individuals were stated as old individuals, in PCA plots of all samples and old and AD samples, the 13th individual was observed at the middle of the plot, and the 18th individual was detected among AD samples meaning that these individuals had similar RNA intensities with those of AD patients, and the similarity was higher for the 18th individual. In addition, the 20th, 27th, and 30th individuals were denoted as individuals with AD. However, the 20th individual was detected among healthy individuals, and 27th and 30th individuals were found at the middle of all PCA plots intermixed with some of the healthy individuals. This indicates the similarity between these individuals' RNA levels with the RNA levels of healthy individuals, and the similarity was higher for the 20th individual. Cluster analysis of samples revealed two clusters formed by healthy and diseased individuals, and the results were consistent with the results of PCA with minor variations. Both scatter plots and dendrogram plots showed that RNA levels in healthy individuals and AD patients were different from each other, so that they formed two separate clusters (See Figures 2.10, 2.11). There were some exceptions as in PCA results. However, for all scatter plots, individuals detected in wrong clusters were only the 18th, 20th, 27th, and 30th individuals. The 28th individual was also found in the wrong cluster only in the scatter plot of young, old, and AD samples. The reason for detecting the 28th individual only in scatter plot of all samples might be the effect of RNA level distributions of young and old samples. Unlike PCA results, the 2nd, 4th, and 13th individuals were detected in the clusters they belonged to, and the 28th individual was detected in the wrong cluster. The separation of samples was not distinct in PCA plots of all samples, so that the 28th individual was not detected by PCA, but detected by cluster analysis. For the dendrogram plots of all samples and the young and AD samples, the 20th, 27th, and 30th individuals were found to be located in wrong clusters as in PCA and scatter plots. However, 27th individual was detected in the correct cluster in dendrogram plot of old and AD samples, while the 13th, 20th, 30th individuals were found in wrong clusters. In contrast to PCA, the 2nd, 4th, and 18th individuals were found to be located in correct clusters. Considering the results obtained by both PCA and cluster analysis, it might be said that there is a high possibility that the 20th, 27th, and 30th individuals were misdiagnosed. Also, the 13th and 18th individuals may have been misdiagnosed even though it is a lower possibility.

Binomial logistic regression analysis was applied to average normalized RNA intensities of individuals for predicting the AD risk. To compare the effect of gene selection, the first three genes, first seven genes, and last nine genes of the selected genes (See Table 2.1) were used for the analysis. It was expected to observe S-shaped curves, and to obtain logit coefficients with P-values lower than 0.05 which indicate a good prediction model. As expected, for analyses where the first three and first seven genes were used, S-shaped curves were obtained and P-values were lower than 0.05. S-shaped curves demonstrated that the RNA levels of AD patients and healthy individuals differed from each other as they were detected at the top, and the bottom of the plot, respectively. However, for the analysis where the last nine genes were used, a line was observed, and the P-values were higher than 0.05. These findings imply that the third model (See Figure 2.12C) is a poor model, and cannot be used for prediction, whereas the other two models (See Figures 2.12A, 2.12B) are statistically significant; therefore, are better models for predicting the AD risk (UCLA: Statistical Consulting Group, n.d.). So, using the genes having the lowest P-values among the selected genes yielded better binary logistic regression models. The prediction codes created based on the first two models were returned more reasonable outputs. However, it must be said that better and more reliable prediction models might be obtained by analyzing a greater number of individuals. In a previous study where binary logistic regression analysis was applied to detect the presence or absence of Eye Glaucoma, it is also recommended to extend the sample size to obtain more accurate and reliable results (Elsalam, 2015).

### 3.2 Conclusion

By analyzing the RNA levels of AD patients and healthy controls, it was shown that gene expression profiles are altered in individuals with AD. The analysis of mean values and standard deviations revealed that the KIAA0922, THRA, and IDH1 genes had similar RNA levels in all samples; whereas the MECR and PTPRD-AS2 genes, and the LOC283440, PRKACB, and LINC01372 genes had distinct RNA levels in young and AD samples, respectively. The LOC283440, PRKACB, and LINC01372 genes were also found in the 80 genes identified (by two sample t-test) as the most significant genes for detecting AD. The first three genes were the LAPTM4B, CXorf56, and RTCA, and their association with AD has been stated in literature as mentioned in the discussion part. The involvement of MECR and PRKACB genes in pathways related to AD development has been also confirmed by previous studies; however, for the remaining genes that had altered RNA levels in young and AD patients, there is no information available regarding their association with AD. In this study, it is suggested that these genes could be somehow linked to AD pathogenesis. Also, it is proposed that the RNA levels of LOC283440, PRKACB, LINC01372 genes and other genes reported in Table 2.1 might be possible biomarkers for Alzheimer’s disease.

By examining the correlation between the RNA levels of all samples, 4 gene pairs (SFR1 & CPPED1, CPPED1 & SGK1, DOCK5 & ATP10B, SFR1 & SGK1) were detected to have significantly correlated RNA intensities in all samples. Even though there is no pathway data available for these genes, these findings indicate their possible involvement in same pathways. It was also found that there was a strong correlation between the RNA levels of 3rd and the 7th AD patients (21st and 25th individuals of the data set) suggesting that AD patients have similar gene expression profiles. Moreover, PCA and cluster analysis showed that healthy controls and AD patients formed different groups based on their RNA levels. The exceptions were mostly the 13th, 18th, 20th, 27th, and 30th individuals of the data set. It is proposed that these individuals might have been misdiagnosed. Finally, binary logistic regression models were created to predict the AD risk of an individual according to the RNA levels of previously determined important genes. The models with S-shaped curves were succesfully obtained by using the first three and first seven genes of the important genes (See Table 2.1). Further research is required to identify more genes related to the development of AD, and to determine their way of involvement. Also, to obtain more reliable results, sample size must be extended, and the number of genes must be increased.

# REFERENCES

2019 Alzheimer's disease facts and figures. (2019). *Alzheimer's & Dementia*. https://doi.org/10.1016/j.jalz.2019.01.010

Alkadhi, K., & Eriksen, J. (2011). The complex and multifactorial nature of Alzheimer's disease. *Current Neuropharmacology*. https://doi.org/10.2174/157015911798376235

Annese, A., Manzari, C., Lionetti, C., Picardi, E., Horner, D. S., Chiara, M., Caratozzolo, M. F., Tullo, A., Fosso, B., Pesole, G., & D'Erchia, A. M. (2018). Whole transcriptome profiling of Late-Onset Alzheimer's Disease patients provides insights into the molecular changes involved in the disease. *Scientific Reports, 8*(1). https://doi.org/10.1038/s41598-018-22701-2

Armstrong, R. A., & Wood, L. (1994). The identification of pathological subtypes of Alzheimer's disease using cluster analysis. *Acta Neuropathologica*, *88*, 60–66. https://doi.org/10.1007/BF00294360

Bartholomew, D. J. (2010). Principal components analysis. In P. Peterson, B. Eva, & B. McGaw (Eds.), *International Encyclopedia of Education* (pp. 374–377). Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.01358-0

Belinky, F., Nativ, N., Stelzer, G., Zimmerman, S., Stein, T. I., Safran, M., & Lancet, D. (2015). PathCards: Multi-source consolidation of human biological pathways. *Database*. https://doi.org/10.1093/database/bav006

Bennett, J. P., & Keeney, P. M. (2018). RNA-Sequencing reveals similarities and differences in gene expression in vulnerable brain tissues of Alzheimer's and Parkinson's diseases. *Journal of Alzheimer's Disease Reports*, *2*(1), 129–137. https://doi.org/10.3233/adr-180072

Berkowitz, C., Mosconi, L., Scheyer, O., Rahman, A., Hristov, H., & Isaacson, R. (2018). Precision medicine for Alzheimer's disease prevention. *Healthcare*. https://doi.org/10.3390/healthcare6030082

Bertram, L. (2016). Next Generation Sequencing in Alzheimer's Disease. *Methods in Molecular Biology (Clifton, N.J.)*. https://doi.org/10.1007/978-1-4939-2627-5_17

Blennow, K., & Zetterberg, H. (2018). Biomarkers for Alzheimer's disease: current status and prospects for the future. *Journal of Internal Medicine*. https://doi.org/10.1111/joim.12816

Boccard, J., & Rudaz, S. (2013). Mass spectrometry metabolomic data handling for biomarker discovery. In H. J. Issaq & T. D. Veenstra (Eds.), *Proteomic and Metabolomic Approaches to Biomarker Discovery* (pp. 425–445). https://doi.org/10.1016/B978-0-12-394446-7.00027-3

Bottero, V., & Potashkin, J. A. (2019). Meta-analysis of gene expression changes in the blood of patients with mild cognitive impairment and alzheimer's disease dementia. *International Journal of Molecular Sciences*, *20*(21). https://doi.org/10.3390/ijms20215403

Carreiras, M. C., Mendes, E., Perry, M. J., Francisco, A. P., & Marco-Contelles, J. (2013). The multifactorial nature of Alzheimer's disease for developing potential therapeutics. *Current Topics in Medicinal Chemistry*, *13*(15), 1745–1770. https://doi.org/10.2174/15680266113139990135

Chaffey, B., & Silmon, A. (2016). Biomarkers in personalized medicine: Discovery and delivery. *Biochemist*. https://doi.org/10.1042/bio03801043

Dana, H., Chalbatani, G. M., Mahmoodzadeh, H., Karimloo, R., Rezaiean, O., Moradzadeh, A., Mehmandoost, N., Moazzen, F., Mazraeh, A., Marmari, V., Ebrahimi, M., Rashno, M. M., Abadi, S. J., & Gharagouzlo, E. (2017). Molecular mechanisms and biological functions of siRNA. *International Journal of Biomedical Science, 13*(2), 48–57.

Defrin, R., Amanzio, M., De Tommaso, M., Dimova, V., Filipovic, S., Finn, D. P., Gimenez-Llort, L., Invitto, S., Jensen-Dahm, C., Lautenbacher, S., Oosterman, J. M., Petrini, L., Pick, C. G., Pickering, G., Vase, L., & Kunz, M. (2015). Experimental pain processing in individuals with cognitive impairment: Current state of the science. *Pain*, *156*(8), 1396–1408. https://doi.org/10.1097/j.pain.0000000000000195

Elsalam, N. M. M. A. (2015). Binary logistic regression to identify the risk factors of Eye Glaucoma. *International Journal of Sciences: Basic and Applied Research (IJSBAR)*.

Gamberger, D., Ženko, B., Mitelpunkt, A., Shachar, N., & Lavrač, N. (2016). Clusters of male and female Alzheimer’s disease patients in the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database. *Brain Informatics*. https://doi.org/10.1007/s40708-016-0035-5

Giri, M., Zhang, M., & Lü, Y. (2016). Genes associated with Alzheimer’s disease: An overview and current status. *Clinical Interventions in Aging*. https://doi.org/10.2147/CIA.S105769

Hamou, A., Simmons, A., Bauer, M., Lewden, B., Zhang, Y., Wahlund, L. O., Westman, E., Pritchard, M., Kloszewska, I., Mecozzi, P., Soininen, H., Tsolaki, M., Vellas, B., Muehlboeck, S., Evans, A., Julin, P., Sjögren, N., Spenger, C., Lovestone, S., & Gwadry-Sridhar, F. (2011). Cluster analysis of MR imaging in Alzheimer’s disease using decision tree refinement. *International Journal of Artificial Intelligence*.

Hattersley, A. T., & McCarthy, M. I. (2005). What makes a good genetic association study? *Lancet*. https://doi.org/10.1016/S0140-6736(05)67531-9

*Introduction to generalized linear models*. (n.d.). Retrieved June 13, 2020, from https://online.stat.psu.edu/stat504/node/216/

Izgi, H. (2017). *Meta analysis of Alzheimer’s disease at the gene expression level* [Middle East Technical University]. http://etd.lib.metu.edu.tr/upload/12620729/index.pdf

Jain, K. K. (2015). Role of biomarkers in personalized medicine. In *Textbook of Personalized Medicine* (pp. 91–97). Humana Press. https://doi.org/https://doi.org/10.1007/978-1-4939-2553-7

Johnson, P., Vandewater, L., Wilson, W., Maruff, P., Savage, G., Graham, P., Macaulay, L. S., Ellis, K. A., Szoeke, C., Martins, R. N., Rowe, C. C., Masters, C. L., Ames, D., & Zhang, P. (2014). Genetic algorithm with logistic regression for prediction of progression to Alzheimer’s disease. *BMC Bioinformatics*, *15*. https://doi.org/10.1186/1471-2105-15-S16-S11

Karch, C. M., Cruchaga, C., & Goate, A. M. (2014). Alzheimer’s disease genetics: From the bench to the clinic. *Neuron*. https://doi.org/10.1016/j.neuron.2014.05.041

Kukar, T. (2019). *A New Approach to Understand Why Defects in the Lysosome Pathway Increase the Risk of Developing Alzheimer’s Disease*. https://www.brightfocus.org/alzheimers-disease/grant/understanding-lysosome-dysfunction-alzheimers-disease-0#:~:text=Details,help degrade and recycle proteins.

Kukurba, K. R., & Montgomery, S. B. (2015). RNA sequencing and analysis. *Cold Spring Harbor Protocols*. https://doi.org/10.1101/pdb.top084970

Landeck, L., Kneip, C., Reischl, J., & Asadullah, K. (2016). Biomarkers and personalized medicine: Current status and further perspectives with special focus on dermatology. *Experimental Dermatology*. https://doi.org/10.1111/exd.12948

Lane, C. A., Hardy, J., & Schott, J. M. (2018). Alzheimer's disease. *European Journal of Neurology*. https://doi.org/10.1111/ene.13439

Lee, R. C. T. (1981). Clustering analysis and its applications. In *Advances in Information Systems Science* (pp. 169–170). Springer. https://doi.org/https://doi.org/10.1007/978-1-4613-9883-7_4

Liang, K.-H. (2013). Transcriptomics. In *Bioinformatics for biomedical science and clinical applications* (pp. 65–66). Woodhead Publishing Series in Biomedicine. https://doi.org/https://doi.org/10.1533/9781908818232.49

Maharzi, N., Parietti, V., Nelson, E., Denti, S., Robledo-Sarmiento, M., Setterblad, N., Parcelier, A., Pla, M., Sigaux, F., Gluckman, J. C., & Canque, B. (2013). Identification of TMEM131L as a Novel Regulator of Thymocyte Proliferation in Humans. *The Journal of Immunology*, *190*(12), 6187–6197. https://doi.org/10.4049/jimmunol.1300400

Mathworks. (n.d.). *Fitting data with generalized linear models*. Retrieved June 13, 2020, from https://www.mathworks.com/help/stats/examples/fitting-data-with-generalized-linear-models.html

Myers, R. H., Montgomery, D. C., Vining, G. G., & Robinson, T. J. (2012). *Generalized linear models: with applications in engineering and the sciences: second edition*. https://doi.org/10.1002/9780470556986

National Institutes of Health. (2019a). *Alzheimer's Disease Genetics Fact Sheet*. National Institute on Aging. https://www.nia.nih.gov/health/alzheimers-disease-genetics-fact-sheet

National Institutes of Health. (2019b). *Causes of Alzheimer's Disease*. National Institute on Aging. https://www.nia.nih.gov/health/what-causes-alzheimers-disease

Pollock, T. B., Mack, J. M., Day, R. J., Isho, N. F., Brown, R. J., Oxford, A. E., Morrison, B. E., Hayden, E. J., & Rohn, T. T. (2019). A fragment of apolipoprotein E4 leads to the downregulation of a CXORF56 homologue, a novel ER-associated protein, and activation of BV2 microglial cells. *Oxidative Medicine and Cellular Longevity*. https://doi.org/10.1155/2019/5123565

Roessner, U., Nahid, A., Chapman, B., Hunter, A., & Bellgard, M. (2011). Metabolomics - The Combination of Analytical Biochemistry, Biology, and Informatics. In *Comprehensive Biotechnology, Second Edition* (pp. 447–459). https://doi.org/10.1016/B978-0-08-088504-9.00052-0

Roy, J., Sarkar, A., Parida, S., Ghosh, Z., & Mallick, B. (2017). Small RNA sequencing revealed dysregulated piRNAs in Alzheimer's disease and their probable role in pathogenesis. *Molecular BioSystems*, *3*. https://doi.org/10.1039/c6mb00699j

Santer, L., Bär, C., & Thum, T. (2019). Circular RNAs: A novel class of functional RNA molecules with a therapeutic perspective. *Molecular Therapy*. https://doi.org/10.1016/j.ymthe.2019.07.001

Schipper, H. M., Maes, O. C., Chertkow, H. M., & Wang, E. (2007). MicroRNA expression in Alzheimer blood mononuclear cells. *Gene Regulation and Systems Biology*. https://doi.org/10.4137/grsb.s361

Seufert, E. B. (2014). Quantitative methods for product management. In *Freemium Economics* (pp. 47–82). https://doi.org/10.1016/b978-0-12-416690-5.00003-8

Snowden, S. G., Ebshiana, A. A., Hye, A., An, Y., Pletnikova, O., O'Brien, R., Troncoso, J., Legido-Quigley, C., & Thambisetty, M. (2017). Association between fatty acid metabolism in the brain and Alzheimer disease neuropathology and cognitive performance: A nontargeted metabolomic study. *PLoS Medicine*. https://doi.org/10.1371/journal.pmed.1002266

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-019-0150-2

Tanzi, R. E., & Bertram, L. (2005). Twenty years of the Alzheimer's disease amyloid hypothesis: A genetic perspective. In *Cell*. https://doi.org/10.1016/j.cell.2005.02.008

Teipel, S. J., Kurth, J., Krause, B., & Grothe, M. J. (2015). The relative importance of imaging markers for the prediction of Alzheimer's disease dementia in mild cognitive impairment - Beyond classical regression. *NeuroImage: Clinical*, *8*, 583–593. https://doi.org/10.1016/j.nicl.2015.05.006

UCLA: Statistical Consulting Group. (n.d.). *Logistic Regression Analysis*. https://stats.idre.ucla.edu/stata/output/logistic-regression-analysis/

Verghese, P. B., Castellano, J. M., & Holtzman, D. M. (2011). Apolipoprotein E in Alzheimer's disease and other neurological disorders. In *The Lancet Neurology*. https://doi.org/10.1016/S1474-4422(10)70325-2

Weller, J., & Budson, A. (2018). Current understanding of Alzheimer's disease diagnosis and treatment. In *F1000Research*. https://doi.org/10.12688/f1000research.14506.1

World Health Organization. (2018). *WHO - The top 10 causes of death*. www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death%0A

Zheng, L. (2012). *Lysosomal involvement in the pathogenesis of Alzheimer's disease* [Linköping University]. http://liu.diva-portal.org/smash/get/diva2:472009/FULLTEXT01.pdf

Ziegler, A., Koch, A., Krockenberger, K., & Großhennig, A. (2012). Personalized medicine using DNA biomarkers: A review. *Human Genetics*. https://doi.org/10.1007/s00439-012-1188-9