Department of Computer Science and Engineering MPhil Thesis

2021

Machine Learning and Bioinformatics Models to Identify the Genetic Link of Neurological Diseases Associated With the Causal Risk Factors

Chowdhury, Utpala Nanda

University of Rajshahi, Rajshahi

Machine Learning and Bioinformatics Models to Identify the Genetic Link of Neurological Diseases Associated With the Causal Risk Factors

SUBMITTED BY

Utpala Nanda Chowdhury

(Academic Year: 2017-2018)

Dissertation submitted to the University to fulfil the requirements for the degree of Master of Philosophy

Department of Computer Science and Engineering
University of Rajshahi
Faculty of Engineering

Principal Supervisor: Dr. Shamim Ahmad

Co-Supervisor: Dr. M. Babul Islam

Co-Supervisor: Dr. Mohammad Ali Moni

June, 2021

Abstract

Neurological diseases (NDs) are causing burgeoning burden to the patients, healthcare sector and the entire society. Hundreds of millions of people are currently affected by various NDs worldwide and the number is increasing very rapidly. The most frequent categories include Alzheimer's disease (AD), Parkinson's disease (PD), epilepsy, Multiple sclerosis (MS), stroke and other cerebrovascular disorders, migraine and other headache, malignant brain tumors such as Glioblastoma multiforme (GBM) etc. Despite numerous research initiatives, preventive and therapeutic options for most of these NDs still remain very limited. Taking AD as an example, fully effective preventive strategies are unavailable till now. Preventive measures usually comprise primary prevention based on risk reducing by identifying influential factors and secondary prevention through early detection and abatement of the disease at initial stage. But the inadequate epidemiological knowledge of AD risk factors and absence of early premortem accurate diagnosis has foiled AD prevention. However, better insight about the co-occurrence of other neurological complications with AD can yield preventive and therapeutic advancement. On the other hand, enhanced understanding about the factors that impacts the response to the treatment could prolong the survival period. For instance, GBM is such an ND with shorten survival period provided the first line treatment include brain surgery followed by chemotherapy and radiotherapy.

In this context, research initiatives to mitigate the information gap regarding how the causative factors affect the cell pathways altered in NDs and their comorbidities can alleviate the disease burden. Availability of high throughput technologies including microarray and next-generation sequencing (NGS) of tissue mRNA to analyse large-scale transcriptomic data have excelled various bioinformatics methodologies as promising tools in biomedical research field. These approaches include differential gene expression analysis, protein-protein interactions (PPIs), gene ontology (GO), metabolic pathway and regulatory factor analysis. Genetic inspection into the transcriptomic data through these tools yields better insight into the molecular pathogenesis of any health condition in junction with its causative factors and related complications. In addition to this, the exponentially increasing amount of accessible biological data has made machine learning techniques as promising means of discovering hidden genetic knowledge. In this thesis, we presented bioinformatics and computational frameworks based on transcriptomic data and machine learning based survival prediction models.

We have done four experiments in this thesis work where three were for AD and another for GBM. In the first and third experiments, we incorporated a series of bioinformatics and computational approaches to inspect the genetic interaction of AD with various causative factors and neurodegenerative consequences respectively through investigating the coexisting differentially expressed genes (DEGs), shared molecular pathways induced by those DEGs and their protein-protein interactome. In the second experiment, we investigated brain and blood transcriptomics and the expression quantitative trait loci (eQTL) data from the Genotype-Tissue Expression (GTEx) project to obtain interesting common pathways in those cells as an attempt for blood based AD biomarkers identification. Finally, we demonstrated some state-of-art machine learning models with high accuracy to detect the signature genes those are significant biomarkers of processes involved in the progression and survival of GBM.

Findings showed significant genetic interactions between AD and the factors of interest signifying how AD incidence and development may be influenced by the risk factors. Putative links between pathological processes in brain tissue and blood cells in AD were also obtained that may allow assessment of AD status using blood samples. Moreover, constitution of functional links between AD and its neurodegenerative comorbidities have been evidenced by which they affect each others development and progression. We furthermore identified putative prognostic biomarker and GBM-chemoresistance related signature genes those may provide new opportunities for therapeutic intervention and can yield a new understanding of the GBM progression and survival.

In sum, our formulated bioinfomatics and computational framework successfully revealed key pathological factors, therapeutic targets, biomarkers and the associated molecular pathways to employ new insight into AD development and progression. Although their utility must be validated through functional studies and clinical investigations, these identifications offered a new window for AD prevention. In addition to this, the prognostic signature genes predicted by the machine learning models could be the candidate for further study and the efficiency of the models could be improved even though they showed satisfactory performance.

DECLARATION

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

ACKNOWLEDGEMENT

First and foremost, praises and thanks to the Almighty God for giving me the opportunity to work with so many wonderful and creative persons throughout my research work to complete it successfully.

I would like to express my deepest gratitude to my principle supervisor Dr. Shamim Ahmad, Professor, Department of Computer Science and Engineering, University of Rajshahi, Bangladesh. His constant encouragement, dedication and keen interest above all his overwhelming attitude to help me had been solely and mainly responsible for the completion of my work. His meticulous scrutiny, scholarly advice and scientific approach have helped me to a great extent to accomplish this task.

It is my pleasure to express my deep sense of thank and gratitude to my co-supervisor Dr. M. Babul Islam, Professor, Department of Electrical and Electronic Engineering, University of Rajshahi, Bangladesh. His prompt inspirations, timely advice with kindness, enthusiasm and dynamism have given me strength and confidence.

I owe a deep sense of gratitude to my another co-supervisor Dr. Mohammad Ali Moni, The University of New South Wales, WHO Collaborating Centre on eHealth, School of Public Health and Community Medicine, Faculty of Medicine, NSW 2052, Australia for giving me the opportunity to do research under his guidance. He has introduced me to the emerging field of bioinformatics and computational biology. He has spent so much of his valuable time to directly teach me methodologies to carry out the research and to present me the research goals and experimental design as clearly as possible. It was a great privilege and honor for me to work under his supervision. I am also very grateful to Dr. Mohammad Ali Moni for giving me opportunity to collaborate with the members of his research group. His group consists some excellent researchers including Professor Valsamma Eapen from The University of New South Wales, Australia, Dr. Julian M. W. Quinn from the Garvan Institute of Medical Research, Australia and Professor Salem A. Alyami from Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia. I am deeply thankful to them for their cordial response in various research problems.

I am extremely thankful to Dr. Bimal Kumar Pramanik, the former Chairman and Professor, Department of Computer Science and Engineering, University of Rajshahi, Bangladesh, the respected current Chairman Subrata Pramanik, Department of Computer Science and Engineering, University of Rajshahi, Bangladesh, and the honorable Dean, Dr. Md. Ekramul Hamid, Faculty of Engineering and Professor, Department of Computer Science and Engineering, University of Rajshahi, Bangladesh for providing me necessary official facilities for my research. I would like to thank all the respected teachers and staffs of this department for their valuable suggestions and co-operations throughout my work.

Finally, I am much more grateful to my family members, especially my parents for their continuous blessings and inspirations, and my wife for her enormous support and best wishes during my research pursuit.

Contents

1 INTRODUCTION **1**
1.1 Background and Motivations . . . 1
1.2 Literature Review . . . 2
1.3 Problem Statement . . . 5
1.3.1 Causative factors of AD . . . 5
1.3.2 Blood based biomarker of AD for early detection . . . 5
1.3.3 Neurodegenerative comorbidities of AD . . . 6
1.3.4 Prolonged survival of GBM . . . 6
1.4 Contribution . . . 6
1.5 Organisation of the Dissertation . . . 7

2 BIOINFORMATICS APPROACH FOR GENETIC INTERACTION IDENTIFICATION **8**
2.1 Introduction . . . 8
2.2 Gene expression analysis . . . 8
2.3 Protein-Protein Interaction Network Analysis . . . 9
2.4 Gene Set Enrichment Analysis . . . 9
2.5 Regulatiory biomolecules identification . . . 10
2.6 Association of eQTL effects between tissues . . . 10
2.7 Semantic Similarity . . . 10

3 MACHINE LEARNING MODELS FOR SURVIVAL ANALYSIS AND SIGNATURE GENE IDENTIFICATION **12**
3.1 Introduction . . . 12
3.2 Cox Proportional Hazards Model . . . 12
3.3 Gradient Boosting Decision Tree . . . 13
3.4 Support Vector Machine . . . 14
3.5 Evaluation measures . . . 15

4 IMPLEMENTATION OF EXPERIMENTS, RESULTS AND DISCUSSION **16**
4.1 Introduction . . . 16
4.2 Experiment Name: Network-based identification of genetic factors in ageing, lifestyle and type 2 diabetes that influence to the progression of Alzheimer's disease . . . 16
4.2.1 Short summary . . . 16
4.2.2 Materials and methods . . . 17
4.2.3 Results . . . 18
4.2.4 Discussion . . . 21

4.3 Experiment Name: Systems biology and bioinformatics approach to identify gene signatures, pathways and therapeutic targets of Alzheimer's disease . . . 25
4.3.1 Short Summary . . . 25
4.3.2 Materials and Methods . . . 26
4.3.3 Results . . . 26
4.3.4 Discussion . . . 33
4.4 Experiment Name: System biology and bioinformatics pipeline to identify comorbidities risk association: neurodegenerative disorder case study 34
4.4.1 Short summary . . . 34
4.4.2 Materials and methods . . . 34
4.4.3 Results . . . 36
4.4.4 Discussion . . . 43
4.5 Experiment Name: Machine learning and bioinformatics models to identify prognostic biomarker and signature genes related to chemoresistance for glioblastoma multiforme . . . 46
4.5.1 Short summary . . . 46
4.5.2 Materials and methods . . . 46
4.5.3 Result . . . 47
4.5.4 Discussion . . . 49

5 CONCLUSION 53
5.1 Introduction . . . 53
5.2 Future Work . . . 53

List of Figures

4.1 Block diagram of the applied analytical approach. 18
4.2 Diseasome networks of the AD with type II diabetes, ageing, sedentary lifestyle, HFD, high dietary red meat, high alcohol consumption, obesity and smoking. AD forms the center of the networks and other red-colored octagon-shaped nodes represent categories of factors and or disease. A link is placed between the disorders and the A) up-regulated genes (sky-blue colored) or b) down-regulated genes (yellow colored) if the alteration of expression of that gene is associated with the specific disorder. 19
4.3 Protein-Protein interaction network for the overlapping DEGs between the AD and other risk factors and diseases. The clusters indicate the risk factors for which the DEGs are dysregulated. 20
4.4 The simplified PPI network depicting the hub genes. The 10 most significant hub genes are marked as red, orange and yellow respectively. . 21
4.5 A pipeline to identify blood cell-expressed Alzheimer's disease biomarkers and pathways that are coordinately expressed in brain tissue. 27
4.6 Identified upregulated DEGs in both brain and blood. A) log fold changes of the common upregulated significant genes in brain and blood and B) Negative logarithmic FDR adjusted p-values of upregulated significant genes common to brain and blood. 28
4.7 Identified downregulated DEGs in both brain and blood. A) log fold changes of the common downregulated significant genes in brain and blood and B) Negative logarithmic FDR adjusted p-values of downregulated significant genes common to brain and blood. 29
4.8 Protein-protein interaction (PPI) network of the AD. The nodes indicate the proteins and the edges indicate the interactions between two proteins. Color indicates MCL analysis clusters of proteins. 32
4.9 The simplified PPI network of the common DEGs between blood and brain AD. The 10 most significant hub proteins are marked as red, orange and yellow respectively. 32
4.10 Pipeline of the analytical approach. 37

4.11 Example DAG of GO terms with GSEA on GSE12685 dataset of AD. The original graph (on the top) and a zoom (on the bottom) are presented. The 5 most significantly enriched GO terms are indicated by the rectangles and the oval shaped nodes represent significant GO terms. The red and orange colors indicate the most significant GO terms. The last two lines inside each node show raw p-value followed by the number of significant genes and the total number of genes annotated to the corresponding GO term for the dataset. 40
4.12 Cluster network with overlapping DEGs between AD and other selected pathologies obtained using the online tool GeneMania. Nodes indicate DEGs and links represent functional associations. The node size indicates the rank of the gene considering its association with other nodes and width of the edges represent the percentile contribution of the connecting nodes to a particular functional association. 41
4.13 Semantic similarity matrix for the differential expressed genes in the five most significant GO terms. The first two letters of each entry represents the selected pathologies (AD-Alzheimer's disease, ALS-Amyotrophic lateral sclerosis, FTD-Frontotemporal dementia, HD-Huntington's disease, LBD-Lewy body disease, MS-Multiple sclerosis, PD-Parkinson's disease). 42
4.14 Semantic similarity matrix for the five most significant GO terms. Entry names are similar as Fig 4.13. 43
4.15 Semantic similarity matrix for DO terms. Legends are: AD-Alzheimer's disease, ALS-Amyotrophic lateral sclerosis, FTD-Frontotemporal dementia, HD-Huntington's disease, LBD-Lewy body disease, MS-Multiple sclerosis, PD-Parkinson's disease. 44
4.16 KEGG pathway enrichment analysis for differentially expressed genes. Each row represents a KEGG pathway associated with the diseases shown in columns. The domination of genes in the pathway indicated by the dimension of the circles and the range of the circles represents the statistical validation for p-value = 0.05. 44
4.17 Block diagram for the multi-stage methodology of the study. 47
4.18 Venn diagram showing overlapping a) up-regullated and b) down-regulated DEGs among three microarray datasets. 48
4.19 A) Protein-protein interaction network for the common DEGs between the three GBM datasets. B) Subnetwork1. C) Subnetwork 2. D) 10 hub genes. 51
4.20 Survival pattern comparison of altered and unaltered expression groups for PHLDA1, IQGAP2, and SPARC. 51
4.21 The frequency of the selected 19 significant signature gene over 20 experiments. 52
4.22 The accuracy, precision and AUC indicators while increasing the number of signature genes in each iteration. 52

5.1 Publication 1 75
5.2 Publication 2 76
5.3 Publication 3 77

List of Tables

4.1 The 10 most significant hub genes. (Ag = Ageing, T2D = Type II Diabetes, Ob = Obesity, Sm = Smoking, HFD = High Fat Diet, RM = Red Meat, AC = high alcohol Consumption, SL = Sedentary Lifestyle.) 21

4.2 Some significant KEGG pathways that are related to the nervous system and common among the AD and other risk factors and diseases. (Ag=Ageing, T2D=Type II Diabetes, Ob=Obesity, Sm=Smoking, HFD=High Fat Diet, RM=Red Meat, AC=Alcohol Consumption, SL=Sedentary Lifestyle.) 22

4.3 Significant GO ontologies that are related to nervous system, and common between the AD and other risk factors and diseases. (Ag=Ageing, T2D=Type II Diabetes, Ob=Obesity, Sm=Smoking, HFD=High Fat Diet, RM=Red Meat, AC=Alcohol Consumption, SL=Sedentary Lifestyle.) 23

4.4 Gene-disease association analysis of DEGs of 7 risk factors and type II diabetes with AD using OMIM and dbGaP databases. 24

4.5 The top 5 significant GO terms for common DEGs between blood and brain tissue in AD. 30

4.6 Significant KEGG pathways for common DEGs between blood and brain tissue in AD. 30

4.7 Significant transcription factors regulating common DEGs between blood and brain tissue in AD. 30

4.8 Significant MicroRNAs regulating common DEGs between blood and brain tissue in AD. 31

4.9 Top 10 significant drug targets identified for common DEGs between blood and brain tissue in AD. 31

4.10 Gene-disease association analysis of the potential biomarkers in blood cells for AD using benchmark databases. 33

4.11 Statistical summary for AD studies. The 3rd, 4th, 5th and 6th columns represent the count for unfiltered genes, the significant DEGs with threshold for p-value, adjusted p-value and logFC (numbers in brackets are for logFC with threshold 2). 7th and 8th columns show the count for unfiltered GO terms and filtered GO terms using Fisher test. 38

4.12 Statistical summary for studies of neurodegenerative comorbid diseases of AD. The 4th, 5th, 6th and 7th columns represent the count for unfiltered genes, the significant DEGs with threshold for p-value, adjusted p-value and logFC (numbers in brackets are for logFC with threshold 2). The 8th and 9th columns show the count for unfiltered GO terms and filtered GO terms using Fisher test. 39

4.13 Summary of findings in the steps of the pipeline for the selected pathologies. . . . 40
4.14 Top 10 significant Biological process GO terms associated with GBM. . . 49
4.15 Significant KEGG pathways associated with GBM and central nervous system. . . . 50
4.16 Significant KEGG pathways associated with GBM and central nervous system. . . . 50

5.1 Concluding remarks regarding the experiments carried out in this thesis. 55

Chapter 1

INTRODUCTION

1.1 Background and Motivations

Neurological diseases (NDs) are the disorders of the central nervous system (CNS) and the peripheral nervous system. The most prevalent NDs include Alzheimer's disease (AD), Parkinson's disease (PD), epilepsy, Multiple sclerosis (MS), stroke and other cerebrovascular disorders, migraine and other headache, malignant brain tumors such as Glioblastoma multiforme (GBM) etc. Usually, most of these NDs fail to be listed as the most lethal diseases. As a result, they have been neglected by the researchers and policy makers worldwide for long periods of time. But, the mortality rate alone can not reflect the seriousness of a disease. Considering only the mortality rate could underestimate the difficulties of a disease that may be non-fatal but cause substantial disabilities. If the disability caused by a disorder is also taken into account, several NDs would emerge as the major causes of suffering. Statistics suggest NDs as the leading cause of disability and the second leading cause of death worldwide in 2015 [1]. The latest Global Burden of Disease (GBD) study included AD and other dementias, PD, MS, epilepsy and headache disorders accounting 3% of the global burden of diseases and ranked these NDs in the top 50 causes of disability-adjusted life-years (DALYs) [2]. Whereas, AD and PD were listed in the top 15 causes having the most considerable raise in burden in the period from 1990 to 2010. The global burden of NDs is estimated to increase substantially in low- and middle-income countries (LMICs)[2]. As the number of the world population as well as the life expectancy are increasing, more people are now falling into the aged groups where NDs are believed to be more prevalent. Despite of having notable impact of NDs on global health, understanding about the epidemiology and knowledge about the causative risk factors and the comorbidities remain minimal, especially in the LMICs. Again, the patients with NDs usually need sufficient financial and social attention because of their cognitive, physio-social and physical disabilities [3]. Thus, the burden caused by the NDs can be expected to be noxious to the poor population. Moreover, these patients often become victim of human rights violations as they get stigmatized and discriminated to access proper medical attention. Therefore, the NDs require significant research attention to improve our knowledge about the disease mechanism, causative factors and consequences, particularly in LMICs.

AD is the leading or emerging ND while considering the most occurrences, the maximum number of patients being untreated, and lack of affordable and effective treatment. AD is the most frequent neurodegenerative disease (NDD) which is considered to be the current primary cause of dementia, causing most of all dementia cases

(60% to 80%). Whereas, dementia is the degradation of cognitive skills and intellectual abilities. The initial symptoms of AD include inability to remember recent events [4]. As the condition progresses, several symptoms including mood swings, lack of motivation, difficulties with language, self-hatred, behavioral abnormality, and disorientation become more prevalent. At later stage, patients become dependant on others, they withdraw themselves from the family and society [5]. Finally, they lead to death trough entirely losing bodily functions [6]. Usually the life expectancy after AD diagnosis may vary from 3 to 9 years. Currently, AD is listed as sixth among all ages and third for older people in the leading death causing diseases in the United States [6]. 5.7 million Americans had AD in 2018, and this number is projected to reach 13.8 million by 2050 [7]. It was a major cause of mortality in 2015, 110,561 deaths from AD were officially recorded in that year in the United States [7]. The main features of AD include cognitive deficiency including memory loss and diminished abilities to carry out simple everyday activities [8], in addition to depression, apathy, hallucinations, delusions and aggression [9]. Significant AD-related features seen in the central nervous system include localized accumulations of beta-amyloid (Aβ) protein in plaques in the extracellular space and tau protein tangles inside neurons. It is not clear whether these are primary causes or pathophysiological responses to AD. But these features can originate over 20 years before AD cognitive symptoms become clearly evident. The pathogenic mechanisms that underlie AD initiation and development are very poorly understood, although a number of genetic and environmental risk factors have been reported to have association with AD [10]. The apolipoprotein E (APOE4) is evidenced to be related to AD throughout the world population [11, 12, 13]. APOE4 is considered as a risk factor for late onset AD, a unfamiliar condition [14]. Genetic studies suggest that mutations involving the amyloid precursor protein (APP) and the presenilin 1 and presenilin 2 protein (PSEN1/2) related genes give rise to plaques [15]. Nevertheless, the inheritance of APP or PSEN1/2 mutants is associated with a high probability for AD development, consistent with an important role for their corresponding proteins [16]. To this day, no disease modifying drugs for AD are available, all the FDA approved drugs only alleviate the symptoms. Most of the clinical trials for AD-therapeutics are Aβ-based and they have failed [17]. The current AD research initiative primarily focus on pathological intervention, causative factor identification for AD prevention, enrichment of knowledge about complications due to AD, formation of early detection and finally therapeutic strategy and drug development [18]. GBM, on the other hand, is the commonest and most lethal tumor of human brain [19]. It accounts 15 percent of all brain tumors still its actual causes and preventing measures are completely unknown [20, 21]. Not only that, the poor prognosis causes death of most patients within one year after first diagnosis [22]. In 2016, where approximately 12,120 incidences of GBM were estimated in the United States, only 5 percent of them survived 5 years [23]. Over the years, considerable amount of research efforts were made but very little progression has been possible for longer survival of GBM [24].

1.2 Literature Review

Neurological diseases cause structural, biochemical or electrical anomalies in the brain, nervous system and spinal cord. These changes provoke a wide range of disabilities with motor and memory function and associated consequences. Most of the AD-related pathology occurs in the brain, just like other NDs [25, 26]. Thus the most accurate diagnosis can be performed by examining the brain tissues through microscopic anatomy, which is known as histology [27]. But, it is immensely challenging to study brain tissue because these tissue samples have to be collected post-mortem with a high degree of cellular heterogeneity. Hence, there is no definite early pre-mortem diagnosis for AD aside from cognitive assessments and the use of brain imaging methods that reveal significant degeneration. The traditional brain imaging approaches include computed tomography (CT) or magnetic resonance imaging (MRI) while the single-photon emission computed tomography (SPECT) or positron emission tomography (PET) are adopted in advanced procedures. All these brain imaging based diagnostic methods are costly and hence are not affordable by all the patients, especially in LMICs. For these cases, usually the diagnosis is performed by examining medical history of the patients and their relatives, and observing their behaviors. The most widely used method to diagnose AD is neuropsychological tests using cognitive test for the evaluation of cognitive degradation [28]. Symptoms of impairments in cognition due to AD emerge only after significant, unchangeable neural degeneration has occurred. Therefore, neuropsychological tests can diagnose AD at later stages and produce very low accuracy at the early stages. Hence there is a need for a simple means to detect AD early, before the indication of cognitive symptoms. An accurate, early diagnostic test for an AD may enable interventions to be more effective and lead the development of new treatments. Identifying robust biomarkers for AD in blood samples could potentially achieve these goals.

In addition to this, better insight about the causative factors, especially the modifiable ones, can yield effective preventive measures to reduce the disease burden. Till date, the pathogenesis of the AD is not clearly understood, but it is clear that both genetic and environmental factors are likely to be significant causes. Numerous research initiatives over the years have confirmed the impact of genetic factors in AD development. Very few cases of APP mutations [29] and relatively larger cases of PSEN1/2 mutations [30] showed links with AD. Alongside, it is also considered that many more regulating genes are yet to be identified [31]. Similarities between the pathological alteration of AD with those of aging has established AD as an accelerated form of aging [32]. Age is the most influential risk factor for AD, along with a sedentary lifestyle. Typically AD develops after the age of 65 years and almost half individuals over 85 years old have AD [33]. Obesity also increases the risk of AD occurrence [34]. Type II diabetes (T2D), hypertension, smoking and dietary fats can also significantly increase the risk of developing AD [35, 36, 37]. Meta-analysis of prospective studies suggests that alcohol consumption in late life yields reduced risk of dementia and hence reduces the risk of AD [38]. Most of these studies are either epidemiological or based on the observation of pathological changes occurred in AD. On the other hand, a number of genetic studies focused on identifying new susceptible genes using traditional linkage-based or candidate-gene-based interconnections [39, 40, 41]. But, as a complex polygenic disorder, genetic association of many of the associated causative factors for AD are yet to be identified.

Meanwhile, symptoms of various neurodegenerative disorders (NDDs) become evident at any point during the course of AD development. Moreover, AD and some other NDDs share similar genetic and environmental risk factors indicating their possible coexistence. Parkinson's disease (PD) is the second-most common NDDs after AD, characterized by the deficiency of striatal dopamine due to the neuronal loss in the substantia nigra, along with deposition of α-synuclein in neurons [42, 43, 44]. Neuronal death and neural dysfunction caused by oxidative stress and mitochondrial DNA (mtDNA) variants are reported to be associated with both AD and PD [45, 46]. Huntington's disease (HD) is usually an inherited and autosomal dominant disorder that causes brain cell damage [47]. Neuropathologic characteristics of PD, HD and AD are evidenced to be consistent that involves neurotoxins in their pathogenesis [48]. Amyotrophic lateral sclerosis (ALS) is a lethal NDD that triggers decay of motor neurons and eventually the control of the motor system is lost [49]. ALS and dementia share genetic sensitivity resulting in their co-occurrences [50]. The TNFα-signaling axis and neuroinflammation, both play a significant role in the pathogenesis of ALS and AD [51]. Spinal Muscular Atrophy (SMA) is mostly an inherited NDD with autosomal recessive nature. Both HD and SMA are entirely monogenic conditions caused by a mutation in the huntingtin gene (HTT) [52] and the SMN1 gene [53] respectively. Lewy Body Disease (LBD) is the primary cause of dementia after AD, particularly in aged groups [54]. The cognitive impairments resulted in both LBD and AD are directly associated with the synaptic loss [55, 56]. α-synuclein is found to have a notable influence in the pathogenesis of LBD and AD [57]. Frontotemporal dementia (FTD) is a focal variety of dementia associated with the continuous deterioration surrounding the prefrontal and anterior temporal cortex [58]. FTD and AD patients show identical executive functions which indicate similar abnormalities in the frontal lobes [59]. Multiple sclerosis (MS) is an inflammatory disease that affects the brain and spinal cord, and results in intellectual trouble [60]. The central nervous system of MS and AD patients exhibit a key contribution of the microglia activation [61]. Therefore, the cognition impairment in AD highly influences the progression and presentation of other NDDs. However, inadequate understanding of AD and its consequences, that means how these NDDs and AD influence each other is unknown. Such co-occurrences can be investigated at a molecular level, for example by identifying genes with altered expression or molecular pathways that are shared by the NDDs and AD. Previously developed data analysis methods for disease comorbidity studies include comoR [62], POGO [63], CytoCom [64], comoRbidity [65] and Comorbidity4j packages [66]. All these methods provide platform for commorbidity analysis within a disease pair, limiting their scope. However, the use of gene expression analyses in the study of comorbidity may offer improved insights into AD disease mechanisms [67]. The availability of huge public transcriptomics resources such as microarray data and bioinformatics tools has enabled us to perform comorbidity analyses, i.e., identify gene pathways that enable two diseases to influence each other [68, 69].

Now-a-days, the exponentially increasing amount of accessible biological data has made machine learning techniques as promising means of discovering hidden genetic knowledge. However, substantial progression in the study of GBM mechanisms have found a variety of factors to influence the patient's prognosis. Age, preoperative quality and extent of tumor resection are amongst the well-known variables. Although multiple indicators have been proposed, the need to clinically test the GBM patients inhibits their effectiveness as predictive biomarkers or something else. To resolve the shortcomings of existing clinical approaches, the design of prediction using imaging features generated from the whole tumor could be a desirable solution which is called multi-parametric MRI. In GBM diagnosis and recovery preparation, it is considered as a highly useful scanning procedure [70]. This technique has some limitations like several tissue samples are gathered from spatially distinct subregions within the same tumor for each patient. The vast majority of biopsy targets are isolated by more than

1 cm. Although 10% of samples are isolated by 5–10 mm, the effects of possible sample duplication can be reduced by low regions of interest (ROI) levels [70]. Another technique named novel radiomic features based on joint intensity matrices (JIMs) are used to detect the survival time of the GBM patient [71]. During carcinogenesis and cancer development, gene microarray and high-throughput sequence technologies allow the study of gene expression profiles. New predictive biomarkers can be identified based on the gene's expression profiles and clinical cases [72]. Chemoresistance (CR), on the other hand, can be acquired by the aberrant expression pattern of several genes [73, 74]. These individual genes as well as their interactions can significantly influence the cancers and CR [75]. Signature genes associated with CR can be successfully identified by taking advantage of several machine learning models [76, 77]. The biological processes and signalling pathways mediated by these genes are well influenced by their genetic interactions with other genes [78]. Thus, the identification of CR-related genes can advance the therapeutic development and drug-response prediction for GBM.

1.3 Problem Statement

As a chronic, most frequent and irremediable dementia, AD is causing a substantially increasing burden to the patients, caregivers and the whole society. This is mostly because, the actual causes of the most AD cases and their consequences are still unknown. On top of that, there is no effective and accurate early diagnosis of AD. Considering these information gap about AD, we incorporated bioinformatics approach to have better insight about AD and its causative factors as well as its neurodegenerative consequences. We also incorporated the expression quantitative trait loci (eQTL) data from the Genotype-Tissue Expression (GTEx) project to identify potential AD biomarkers and molecular pathways of the brain which behave similarly in blood. Besides this, as an attempt to mitigate the research gap regarding poor prognosis and paucity of knowledge about CR related genes for GBM, we utilized machine learning methodologies to identify prognostic markers and signature genes for GBM-CR to advance towards prolonged survival.

1.3.1 Causative factors of AD

Prevention of AD at early stage requires proper insight about the disease mechanism and the factors influencing its occurrence. Although a number of strong causative factors are known, the actual causes of the most AD cases are poorly understood. Understanding how these risk factors affect cell pathways that are altered in AD could identify important causal pathways that could be targeted by therapeutics. In this thesis, we proposed a network-based quantitative framework to investigate these factors-AD association. Here, we considered aging, type II diabetes and several lifestyle factors that includes smoking, excessive alcohol consumption, obesity, sedentary lifestyle, high dietary fat and high dietary red meat as influential causative factors of AD.

1.3.2 Blood based biomarker of AD for early detection

Blood based biomarker can make possible the effective and accurate diagnosis of AD at early stage. Since most genes are regulated differently in different tissues, we need to identify genes that are similarly expressed in blood and brain cells. If this proves possible, it could enable blood cell transcript profiles to become a window into some of the pathological changes affecting the brain. In this thesis, we investigated common transcripts (using microarray data) of brain and blood for altered in AD affected individuals, to find biomarkers that are expressed under similar genetic control in both cells using eQTL data. To identify potential biomarkers and molecular pathways of the brain which behave similarly in blood, we analyzed human genomic and transcriptomic data and examined the involvement of the genes on pathogenic processes using curated gold benchmark databases for AD. The primary goal is to discover biomarker signatures in the blood that could help the diagnosis of AD at an early stage and so add valuable information to our understanding of the molecular mechanisms that underlie AD.

1.3.3 Neurodegenerative comorbidities of AD

A wide spectrum of comorbidities, including other neurodegenerative diseases, are frequently associated with AD. How AD interacts with those comorbidities can be examined by analysing gene expression patterns in affected tissues using bioinformatics tools. In this thesis, we surveyed public data repositories for available gene expression data on tissue from AD subjects and from people affected by neurodegenerative diseases that are often found as comorbidities of AD. We then utilized large set of gene expression data, cell-related data and other public resources through an analytical process to identify functional disease links. This process incorporated gene set enrichment analysis and utilized semantic similarity to give proximity measures. We identified genes with abnormal expressions that were common to AD and its comorbidities, as well as shared gene ontology terms and molecular pathways.

1.3.4 Prolonged survival of GBM

It is essential to identify the causal genetic targets associated with GBM survival. Plenty of publicly accessible gene expression and clinical data for GBM patients from the Gene expression omnibus (GEO) and The Cancer Genome Atlas (TCGA) dataset can allow us to study patient fatality prediction and thus to identify new GBM biomarkers. In this thesis, we applied machine learning and bioinformatics approach to identify the altered genes associated with the GBM comparing with normal mRNA expression data from the brain tissues. Besides this, identification of the individual signature genes associated with the reduced responsiveness to the chemotherapy while treating GBM can help advancing therapeutic development. In this thesis, we also analyzed the TCGA data of GBM through machine learning models to identify the signature genes that influence acquiring chemoresistance in GBM patients.

1.4 Contribution

In this thesis, at first, we formulated bioinformatis methodologies using network-based approach to identify significant genetic interactions between AD and other related health factors that may be impacted. Findings suggests significant contribution of several causal factors to AD development and also reported some signature genes for therapeutic intervention. Secondly, we employed a transcriptional analysis of AD affected blood and brain tissues, and integrated them with cis-eQTL data. Here, we identified new putative links between pathological processes in brain tissue and blood cells in AD that may allow assessment of AD status using blood samples. After that, we developed a semantic similarity based methodological pipeline to highlight the factors and pathways that may constitute functional links between AD and its neurodegenerative comorbidities. We then constructed a Cox Proportional Hazard regression-based prognostic model to determine significant biomarkers that are involved in GBM survival. Three signature genes were resulted to have significant dominance on the survival of GBM patients. Finally, we incorporated gradient boosting decision tree (GBDT) and support vector machine (SVM) to identify CR related signature genes for GBM. Nineteen candidate signature genes showing notable correlation with GBM-CR were obtained by the machine learning models.

1.5 Organisation of the Dissertation

The remaining of this thesis is divided into another four chapters. Chapter 2 will introduce the bioinformatics methodologies that can be used to investigate the genetic association between different health conditions and their causative factors. Chapter 3 will highlight the machine learning models that can be incorporated for survival prediction and signature gene identification. Chapter 4 will discuss the formation and analysis of our experiments. We have implemented four experiments to identify the genetic link of different causal factors for neurological diseases. The first experiment has been done to reveal candidate biomarkers that may enhance understanding of mechanisms underlying AD and their link to the risk factors. In the later experiment, we investigated brain and blood transcriptomics and eQTL data to identify interesting common pathways in those cells. The third experiment aimed to analyse the transcriptome of AD and other neurodegenerative diseases that are common comorbidities. The final experiment has been carried out to identify signature genes that may have significant influence on GBM survival and GBM-CR. Finally, conclusions, future directions, and recommendations are presented in Chapter 5.

Chapter 2

BIOINFORMATICS APPROACH FOR GENETIC INTERACTION IDENTIFICATION

2.1 Introduction

Availability of high throughput technologies to analyse large-scale transcriptomic data have excelled various bioinformatics methodologies as promising tools in biomedical research field [79]. The key genetic factors associated with susceptibility to complex diseases may be unravelled by genome-wide association studies, indeed the usefulness of this approach has been proven empirically [80]. This type of molecular association analyzes, which includes differential gene expression determination, protein-protein interactions (PPIs), gene ontology (GO), metabolic pathway and regulatory factor analyzes can ascribe gene activity-based relationships among the disease conditions [81].

2.2 Gene expression analysis

Analyzing oligonucleotide microarray data for gene expression is an effective approach to identify new molecular determinants of human diseases. In this study, we used this methodology along with global transcriptome analysis to investigate the gene expression profiles of the AD with 8 risk factors. To mitigate the problems involving messenger RNA (mRNA) data comparison using different platforms and experimental set-ups, we normalized each gene expression data for each disease using the Z-score (or zero mean) transformation for both disease and control state. Each sample of the gene expression matrix was normalized using mean and standard deviation. The expression value of gene i in sample j represented by g_{ij} was transformed into Z_{ij} by computing

$$Z_{ij} = \frac{g_{ij} - mean(g_i)}{SD(g_i)} \tag{2.1}$$

where SD is the standard deviation. Comparing values of gene expression for various samples and diseases are made possible by this transformation.

Data were transformed using log_2 and unpaired student t-test was performed to find the best candidate genes (with the greatest difference in expression) for identifying

DEGs by using threshold values. A threshold for p-value and absolute base two log fold change (logFC) values were set to at most 0.05 and at least 1.0 respectively.

D is a specific set of diseases and G is a set of dysregulated genes, gene-disease associations attempt to find whether gene $g \in G$ is associated with disease $d \in D$. If G_i and G_j, the sets of significant dysregulated genes associated with diseases D_i and D_j respectively, then the number of shared dysregulated genes (n^g_{ij}) associated with both diseases D_i and D_j is as follows [82]:

$$n^g_{ij} = N(G_i \cap G_j) \tag{2.2}$$

Here $N(G_i)$ is the number of dysregulated genes associated with disease D_i. The common neighbours are obtained based on the Jaccard Coefficient method, where the edge prediction score for the node pair based on the similarity among them is measured as:

$$E(i, j) = \frac{N(G_i \cap G_j)}{N(G_i \cup G_j)} \tag{2.3}$$

where G is the set of nodes and E is the set of all edges.

2.3 Protein-Protein Interaction Network Analysis

Proteins exhibit physical contacts with each other in a cell or in a living organism indicating some biochemical events, typically functions as some molecular processes within a cell, and thereby forms a protein-protein interaction (PPI) network [83]. Such PPI network for the DEGs can be constructed using STRING. STRING provides knowledge base about known and estimated PPIs that comprises both physical and functional interactions, where nodes represents genes and edges indicates interconnection between them. At present, this database includes 24,584,628 proteins from 5,090 organisms [84]. Proteins with different network characteristics such as having high-degree of interactions, may have significant role in the cellular responses to a special physiological stimulus. Such highly interconnected nodes of the network, known as *hub* genes, can be identified using *cytoHubba* plugin of Cytoscape software [85] with the Degree topological algorithm [86]. These hub genes produce a highly dense module inside the interactome that could be of importance in effective drug discovery. We have extracted such highly concentrated modules by analysing the PPI network by another Cytoscape plugin, namely Molecular Complex Detection (MCODE) [87].

2.4 Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) is the procedure of identifying differentially expressed genes (DEGs) in a large set of genes, that may be correlated with disease phenotypes [88]. It uses a set of statistical methods to group genes considering the commonality in their expression level, biological process or chromosomal position. This is done by comparing the expression pattern in disease condition and healthy state. These genes may be acquired using DNA microarray or next-generation sequencing (NGS). The genes having a decisive level of expression are picked up as DEGs (both over and under-expressed).

Gene set enrichment analysis (GSEA) for a set of genes identifies their significant involvement in a certain molecular pathway or functional category to yield knowledge about the biological corollary, position on chromosome, or regulation they share [88]. Such functional categories are defined by gene ontology (GO) terms that are further categorised as biological process (BP), cellular component (CC) and molecular functions (MF) [89]. Similarly, the Kyoto Encyclopedia of Genes and Genomes (KEGG) database provides functional knowledge regarding the cellular processes for analysing signalling pathway [90].

2.5 Regulatiory biomolecules identification

Transcription factors (TF) interacts with genes to regulate the gene expression level by turning it on and off at the transcription level. We utilized JASPAR [91] and TRANSFAC [92] databases via Enrichr to spot the TF-target gene interactions with the target DEGs. Similarly, we used the miRTarBase database by Enrichr to collect the microRNA(miRNA)-gene interaction that controls the target gene at the post-transcriptional stage [93]. Moreover, we used the Drug Signatures Database (DSigDB)[94] to analyze drug-target over-representation on the same platform for candidate drug identification.

2.6 Association of eQTL effects between tissues

Let $\tilde{x}$ be the estimated effect at the top-linked cis-eQTL for a gene between tissues (e.g. blood and brain). We can calculate $\tilde{x}$ as

$$\tilde{x} = x + \epsilon \tag{2.4}$$

where x is the actual effect and ϵ is the estimated error. We assume that x and ϵ are random variables across the genes between tissues, i.e., $x \sim N(0, var(x))$ and $\epsilon \sim N(0, var(e))$. The method is derived from [95] based on eQTL data.

2.7 Semantic Similarity

Semantic similarity is a measure of similarity between terms (DEGs, GO, DO) using ontologies by estimating a topological closeness [96]. This method uses directed acyclic graphs (DAGs) to compute the information contented by each terms considering statistical annotations. The exact position of these terms in the DAG and the connection with their predecessor terms determines the semantic measure. An ontology term T can be denoted by the DAGs $DAG_T = (T, A_T, E_T)$, where A_T is a set of ancestor terms of T and E_T is a set of edges connecting the terms in DAG_T that represent the semantic relation. At first, the semantic measure of each term is represented numerically as,

$$\begin{cases} \mathrm{S}_T(T) = 1 & \text{t=T} \\ \mathrm{S}_T(t) = max\{w_e * S_T(t^{'}) | t^{'} \in \text{ decendants of } (t)\} & \text{t} \neq T \end{cases} \tag{2.5}$$

Here t is a general term, $t^{'}$ a descendant term and w_e the semantic participation of t with $t^{'}$. The inclusive semantic measure for T is

$$SM(T) = \sum_{t \in A_T} S_T(t) \tag{2.6}$$

Now, if $DAG_X = (X, A_X, E_X)$ and $DAG_Y = (Y, A_Y, E_Y)$ are two terms X and Y respectively, then their semantic similarity is

$$sem\_sim(X,Y) = \frac{\sum_{t \in T_X \cap T_Y} [S_X(t) + S_Y(t)]}{SM(X) + SM(Y)} \tag{2.7}$$

Given two sets of terms $T_1 = \{t_{11}, t_{12}, ...t_{1l}\}$ and $T_2 = \{t_{21}, t_{22}, ...t_{2m}\}$ having lengths l and m respectively, the semantic similarity the term sets T_1 and T_2 is

$$sem\_sim_{BMA}(T_1, T_2) = \frac{\sum_{i=1}^{l} max_{1 \leq j \leq m} sem\_sim(t_{1i}, t_{2j}) + \sum_{j=1}^{m} max_{1 \leq i \leq l} sem\_sim(t_{1i}, t_{2j})}{l + m} \tag{2.8}$$

with i, j indices on T_1, T_2 terms.

Chapter 3

MACHINE LEARNING MODELS FOR SURVIVAL ANALYSIS AND SIGNATURE GENE IDENTIFICATION

3.1 Introduction

Survival analysis explores the advantage of machine learning techniques to predict the occurrence of a disease or death of a patient due to the disease. Machine learning models that can predict the survival of a patient can be crucial in understanding the underlying mechanism of a disease progression and mortality. The Cox proportional hazards model [97] is the most widely used approach for survival prediction, especially in cancer prognosis. Again, the microarray and next generation sequencing (RNA-seq) studies have exhibited that gene expression data can be associated with the survival risk of a patient [98]. This provides an opportunity to utilize the gene expression data along with clinical features for survival prediction [99, 100]. For CR-related signature gene identification, we can consider the task as a classical classification problem where the expression of individual genes can be fed to any machine learning model as features. Among various machine learning algorithms, the GBDT, an ensemble learning technique, has been widely applied in genetic research and bioinformatics tasks [101, 102]. Besides this, SVM has also been widely used models applied on high dimensional gene expression data [103, 104].

3.2 Cox Proportional Hazards Model

For survival analysis, we estimated the expected time period until an event of interest occurs, such as a death in cancer. We applied the product-limit (PL) estimator as a survival predictor for this purpose. A log-rank test was used to find out whether the survival function of patients with and without transformed gene expression has a statistically significant contrast or not. A Cox Proportional Hazards (CoxPH) regression model was then constructed for the determination of significant genes. Finally, we conducted functional analyses for the identified significant genes. Each gene was categorised as *altered* or *normal* classes by comparing the gene expression z-score with a threshold value ($z < 2$). We carried out the survival analysis in three steps. This included time to the event, the status, i.e., which patients have to be kept for the analysis, and the event, namely which patients die after initial pathological diagnosis (IPD). We formulated the censoring criterion for the subject included in the investigation, then the time of incidence for an event. The PL estimator is determined as:

$$\widehat{S}(t) = \prod_{i=1}^{k} \left(1 - \frac{d_i}{n_i}\right) \tag{3.1}$$

where $\widehat{S}(t)$ is the predicted survival function at a given time t, d_i is the count for deaths that occur at $t_i < t$, and n_i is the count for patients that remain at t_i. For instance, we implemented log-rank test to identify the statistical significance of genes whose status of being altered or not have a close relationship with patient survival. The null hypothesis formulated to test this is as follows:

$H_0 : S_{altered}(t) = S_{not\ altered}(t)$

$H_A : S_{altered}(t) \neq S_{not\ altered}(t)$

The CoxPH regression model is applied to correlate one or more risk factors or triggers or predictors, usually termed as covariates, to survival time by a hazard function which represents a risk of failure in the model. Therefore, we fit a CoxPH regression model on the combination of all the selected clinical factors and the DEGs as given below:

$$\begin{aligned} f(t|X) &= f_0(t)exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n) \\ &= f_0(t)exp(\beta^T X) \end{aligned} \tag{3.2}$$

Here $f(t|X)$ is the conditional hazard function at time t given that the vector X is a set of covariates information, $f_0(t)$ is the baseline hazard function at time t, and the vector β represents the regression coefficients to X. We used the vector β for calculating the hazard ratio (HR) from the modelled CoxPH to find whether a particular covariate has an influence on patient survival or not. The HR for a covariate x_r is estimated by the exponential function $exp(\beta_r)$.

3.3 Gradient Boosting Decision Tree

GBDT is an ensemble technique based on the classification and regression tree (CART), a variation of decision trees. Here, gradient boosting is used to reduce the residual and consequently the performance is enhanced. For a given set of n known training samples $X = (x_1, y_1), (x_2, y_2), ...(x_n, y_n)$, the machine learning algorithms usually finds a function $F(x)$ that estimates the class label y from a set of input features x. The algorithm formulates the approximation function $\hat{F}$ to minimize the loss function $L(y, F(x))$ as:

$$\hat{F} = \arg\min_{F} \mathbb{E}_{x,y}[L(y, F(x))] \tag{3.3}$$

The gradient boosting approximates a weighted sum of M base learners as:

$$\hat{F}(x) = \sum_{i=1}^{M} w_i h_i(x) \tag{3.4}$$

GBDT starts the classification task by constructing an onset model of a constant function $F_0(x)$ and expanding it gradually:

$$F_0(x) = \arg\min_w \sum_{i=1}^{n} L(y_i, w), \tag{3.5}$$

$$F_m(x) = F_{m-1}(x) + w_m h_m(x) \tag{3.6}$$

The gradient descent function updates the models in each iteration by selecting the parameter w_m using the following equation ensuring that the loss function of the previous model $L(y_i, F_{m-1}(x_i))$ decreases.

$$w_m = \arg\min_w \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + wh_m(x_i)\right) \tag{3.7}$$

GBDT uses CARTs as its base learner $h_m(x)$ having J_m leaves that can be represented as:

$$h_m(x) = \sum_{j=1}^{J_m} b_{jm} \mathbf{1}_{R_{jm}}(x) \tag{3.8}$$

In GBDT, the average feature importance of jth feature for all CART (T) is defined using the Gini index as

$$FI_j(T) = \sum_{k=1}^{L} \Delta_{Gini} I(v_t = j) \tag{3.9}$$

Over N number of experiments, the frequency of the jth feature can be formulated as

$$f_j = \sum_{k=1}^{N} sign(FI_j^{(k)}) \tag{3.10}$$

where $sign(.)$ has the functional representation

$$sign(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \end{cases} \tag{3.11}$$

3.4 Support Vector Machine

Support vector machine (SVM) is one of the most widely adopted supervised machine learning algorithm in almost every domain for both binary and multiclass classification. It is also used in regression task as well as classification task. For any given set of n known training samples $X = (x_1, y_1), (x_2, y_2), ...(x_n, y_n)$, SVM finds the hyper-plane to predict the class label of a particular sample x using the decision function

$$f(x) = sign((w.x) + b) \tag{3.12}$$

SVM learning can be formulated as a constrained optimization problem for w and ξ as

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C \sum_{k=1}^{n} \xi_k, k = 1, 2, ..., n \tag{3.13}$$

$$s.t. \begin{cases} y_k(w.x_k + b) \geq 1 - \xi_k, k = 1, 2, ..., n \\ \epsilon_i \geq 0 \end{cases} \tag{3.14}$$

The Lagrangian formulation of this SVM problem is

$$\min_{a} \frac{1}{2} \sum_{k=1}^{n} \sum_{l=1}^{n} a_k a_l y_k y_l (x_k.x_l) - \sum_{k=1}^{n} a_k \tag{3.15}$$

$$s.t. \begin{cases} \sum_{k=1}^{n} a_k y_k = 0 \\ 0 \leq a_k \leq C, k = 1, 2, ..., n \end{cases} \tag{3.16}$$

Instead of optimizing over w, b, subject to constraints involving a's, the Lagrangian dual problem maximizes the dual variables over a subject to w and b. The solution satisfies:

$$\begin{cases} w = \sum_{k=1}^{n} a_k y_k x_k \\ b = y_l - \sum_{k=1}^{n} y_k a_k (x_k.x_l), 0 < a_k < C \end{cases} \tag{3.17}$$

3.5 Evaluation measures

Various evaluation matrices are used to asses the performance of any machine learning model. Among them accuracy, precision and area under curve (AUC) are well accepted indicators in genetic research. If a sample actually belongs to class C_1 and it is predicted as class C_1 by the model then it labeled as true positive (TP); however, if the model predicts it as class C_2 then the situation is false positive (FP). On the other hand, if the sample belongs to class C_2 and the model classifies it as class C_2, then it is true negative (TN), and false negative (FN) when it is classified as class C_1. Then the evaluation measures can be defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.18}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.19}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.20}$$

$$AUC = \int_{x=0}^{1} Recall(Specificity^{-1}(x))dx \tag{3.21}$$

$$whereSpecificity = \frac{TN}{TN + FP} \tag{3.22}$$

Chapter 4

IMPLEMENTATION OF EXPERIMENTS, RESULTS AND DISCUSSION

4.1 Introduction

In order to reveal the genetic association of various causal factors with the neurological diseases, we carried out four experiments focusing on AD and GBM. The first experiment investigates the genetic linkage of ageing, type 2 diabetes and various lifestyle factors on AD development. The second experiment identifies potential biomarkers and molecular pathways of the brain which behave similarly in blood by analysing human genomic and transcriptomic data. The third experiment demonstrate how AD and other NDDs impact each other at the molecular level through a series of bioinformatics and computational approaches. Whereas, the final experiment incorporates machine learning models to identify prognostic markers and CR-related signature genes to extend survival in GBM.

4.2 Experiment Name: Network-based identification of genetic factors in ageing, lifestyle and type 2 diabetes that influence to the progression of Alzheimer's disease

4.2.1 Short summary

The pathogenesis of the AD is not clearly understood, but it is clear that both genetic and environmental factors are likely to be significant causes. In this experiment, we have used a network-based analysis to determine the genes that mediate influences of associated risk factors and disorders for AD progression, including studies of gene expression profiling, PPI sub-network, gene ontologies and molecular pathways. An extensive study regarding phylogenetic and pathway analysis was therefore conducted to reveal such genetic associations in AD. The significance of these genes and pathways in AD processes were also further validated with gold benchmarking datasets including Online Mendelian Inheritance in Man (OMIM) and dbGaP gene-disease associations databases.

4.2.2 Materials and methods

Data

We analyzed gene expression microarray datasets to identify the alterations of gene expression with AD. All the datasets used in this study were collected from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus database (`https://www.ncbi.nlm.nih.gov/geo/`). Ten different datasets were used for this study, which had, namely those with accession numbers GSE1297, GSE23343, GSE15524, GSE25941, GSE1786, GSE68231, GSE25220, GSE52553, and GSE4806 [105, 106, 107, 108, 109, 110, 111, 112, 113, 114]. The AD dataset (GSE1297) was obtained by gene expression profiling of hippocampal tissues on 31 Affymetrix U133A microarrays from nine control subjects and 22 AD patients that varied in disease severity. The T2D dataset (GSE23343) contains gene expression data using Affymetrix U133A human gene microarrays from liver biopsies in 10 individuals with and 7 without T2D. The obesity dataset (GSE15524) was generated using Codelink Uniset 20K human gene arrays of subcutaneous and omental adipose tissue analyzed by expression profiling 28 tissue samples of obese and lean individuals. The advanced age dataset (GSE25941) consists of data obtained from Affymetrix Human Genome U133 Plus 2.0 array analysis of skeletal muscle transcriptomes of 28 different subjects of older (78 ± 1 years old) and younger (25 ± 1 years old) individuals. The sedentary lifestyle dataset (GSE1786) was obtained by expression profiling by Affymetrix Human Genome U133A arrays of from the quadriceps (vastus lateralis) muscle samples were taken using needle biopsies from sedentary 67 ± 2.5 year old males before and after 3 months of aerobic active training. The high-fat diet (HFD) dataset (GSE68231) is the expression data from Affymetrix Human Genome U133 Plus 2.0 array analysis of human skeletal muscle from 10 people before and after 3 days of high fat diet; these people had previously been identified as highly responsive to the diet, detected by their rapid accumulation of intramyocellular lipid (IMCL). Similarly the red meat dietary intervention dataset (GSE25220) is an Agilent-014850 whole human genome microarray data from human colon biopsies of 22 inflammatory bowel patients before and after participating in a high red-meat dietary intervention. The alcohol consumption dataset (GSE52553) is an Affymetrix human gene expression array data of Lymphoblastoid cells from 21 alcoholics and 21 control subjects. The smoking dataset (GSE4806) is an Affymetrix Human Genome U133A 2.0 array gene expression profiles of T-lymphocytes from 3 smokers and 3 non-smokers.

Analytical Approach

We applied an analytical approach to identify the genetic links between the eight risk factors including T2D and the AD by employing the selected microarray datasets as shown in Figure 4.1. This quantitative approach first determines DEGs for all the risk factors and T2D, and then identifies the overlapping DEGs with the AD study. Further, these common DEGs are used to seek the signaling and ontological pathways and protein-protein interaction (PPI) networks shared by risk factors and the AD. Finally, the proposed approach uses the two gold benchmark databases including Online Mendelian Inheritance in Man (OMIM) and dbGAP to validate the genes and pathways identified in our study as showing possible diseasome risk.

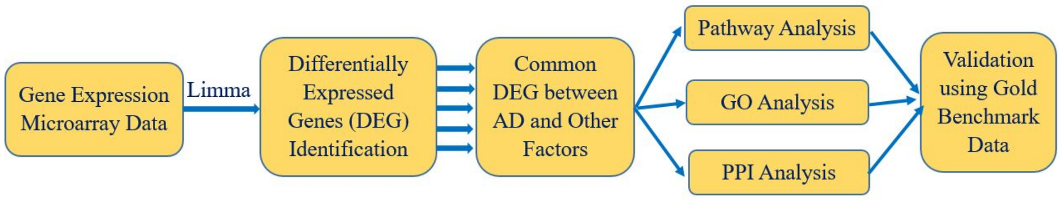

Figure 4.1: Block diagram of the applied analytical approach.

4.2.3 Results

Identification of the Differentially Expressed Genes (DEGs)

Differential gene expression analysis identified 484 genes to be differentially expressed in AD patients compared to healthy subjects where 336 genes were up-regulated and 148 genes were down-regulated. In order to investigate the relationship of the AD transcriptome with that of the risk factors, we next performed several steps of statistical analysis for mRNA microarray data for each risk factor. Thus, we selected the most significant over and under regulated genes for each risk factor and disease. Our analysis identified a large number of DEGs, namely 958 genes in advanced ageing, 1405 in high alcohol consumption, 739 in high fat diet (HFD), 381 in obesity, 482 in high dietary red meat, 800 in sedentary lifestyle, 400 in smoking and 1438 in diabetes type II datasets. The over- and under- expressed genes identified as in common between AD and other risk factors and diseases were also detected through a cross-comparative analysis. The common DEGs of each risk factor are considered to influence to the disease progression. The findings demonstrated that AD shares a total of 35, 34, 18, 15, 13, 10, 8 and 4 significant DEGs with, T2D, alcoholism, sedentary lifestyle, ageing, HFD, obesity, smoking, and high dietary red meat datasets respectively. Two diseasome associations networks centered on AD data were built using Cytoscape to identify statistically significant associations among these risk factors and diseases. Networks shown in Figure-4.2 interpret the association among up- regulated genes and down- regulated genes. Notably, 3 significant genes, HLA-DRB4, IGH and IGHA2 are commonly up-regulated in AD, T2D, and alcoholism datasets; 2 significant genes IGHD and IGHG1, were commonly up regulated among the AD, T2D, alcoholism and sedentary lifestyle datasets. It is noteworthy that a relatively higher number of DEGs were identified as in common between the AD and T2D datasets, whereas the AD and high dietary red meat shared only 4 DEGs. Hence, both T2D and alcoholism are found to have the most influential contribution to AD as they share most DEGs with AD. However, both sedentary lifestyle and ageing tend to have more impact on AD compared to other three factors.

Protein-protein interaction (PPI) analysis to identify common sub-networks

The PPI network was constructed using all the distinct 108 (from total 144) DEGs that were identified as in common between the AD and other risk factors and disease datasets (Figure-4.3). Each node in the network represents a protein and an edge indicates the interaction between two proteins. The network was also grouped into 9 clusters representing risk factors and diseases to depict the protein links. It is notable that KCNJ5 protein belongs to the highest number (3) of clusters indicating that it

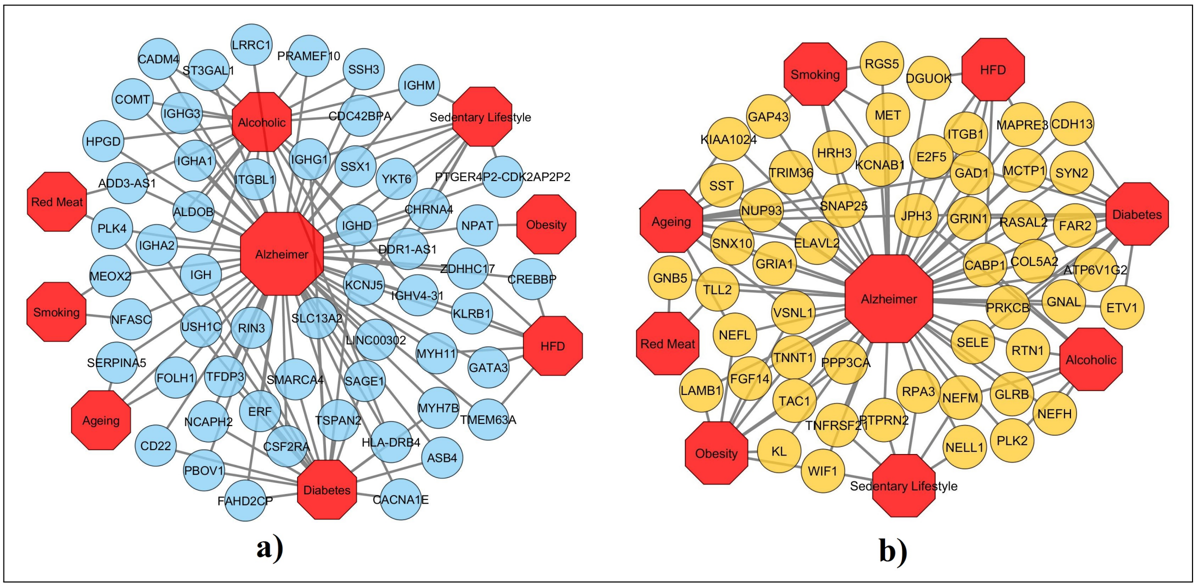

Figure 4.2: Diseasome networks of the AD with type II diabetes, ageing, sedentary lifestyle, HFD, high dietary red meat, high alcohol consumption, obesity and smoking. AD forms the center of the networks and other red-colored octagon-shaped nodes represent categories of factors and or disease. A link is placed between the disorders and the A) up-regulated genes (sky-blue colored) or b) down-regulated genes (yellow colored) if the alteration of expression of that gene is associated with the specific disorder.

is the gene most commonly found among the AD, alcoholism, HFD and sedentary lifestyle datasets and interacts with other proteins from different clusters. In addition the protein products of PLK4, E2F5, GAD1, VSNL1, and CABP1 belong to two clusters each and interact with other proteins in the network. For topological analysis, a simplified PPI network was constructed using Cyto-Hubba plugin to show 10 most significant hub proteins (Figure-4.4), which are CREBBP, PRKCB, ITGB1, GAD1, GNB5, PPP3CA, CABP1, SMARCA4, SNAP25 and GRIA1. Table-4.1 summarises the descriptions and associated risk factors for these proteins. Notably, 4 hub genes among the 10 are dysregulated due to HFD and 3 genes are dysregulated due to ageing. But, sedentary lifestyle is found to have no contribution to the hub genes.

Functional enrichment of DEG sets

In order to identify the molecular pathways associated with the AD and predicted links to the affected pathways, we performed pathway analysis on all the DEGs that were common among the AD and other risk factors and diseases using the KEGG pathway database (`http://www.genome.jp/kegg/pathway.html`) and the web-based gene set enrichment analysis tool EnrichR [115]. A total of 115 pathways were found to be over-represented among several groups. Notably, nine significant pathways that are related to the nervous system were found which are Long-term potentiation (hsa04720), Synaptic vesicle cycle (hsa04721), Retrograde endocannabinoid signaling (hsa04723), Glutamatergic synapse (hsa04724), Cholinergic synapse (hsa04725), Serotonergic synapse (hsa04726), GABAergic synapse (hsa04727), Dopaminergic synapse (hsa04728), and Long-term depression (hsa04730). These pathways along with some other common pathways found are shown in Table 4.2.

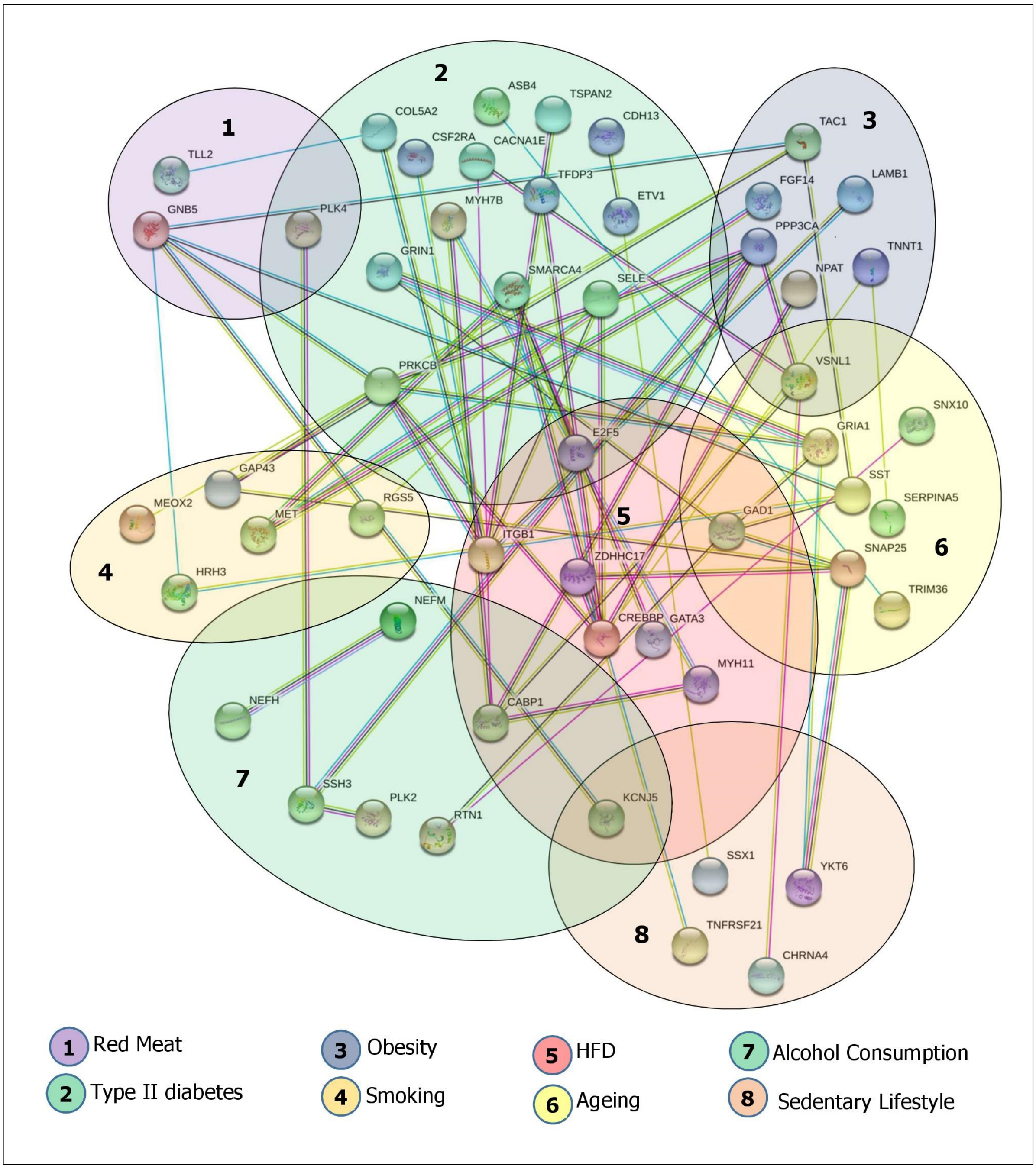

Figure 4.3: Protein-Protein interaction network for the overlapping DEGs between the AD and other risk factors and diseases. The clusters indicate the risk factors for which the DEGs are dysregulated.

We identified over-represented ontological groups by performing gene biological process ontology enrichment analysis using EnrichR on the commonly dysregulated genes between the AD and AD risk factors. A total of 215 significant gene ontology groups were observed; these included peripheral nervous system neuron development (GO:0048935), neurotransmitter transport (GO:0006836), neuromuscular synaptic transmission (GO:0007274), peripheral nervous system development (GO:0007422), negative regulation of neurological system process (GO:0031645), regulation of neurotransmitter secretion (GO:0046928), regulation of neuronal synaptic plasticity (GO:0048168),

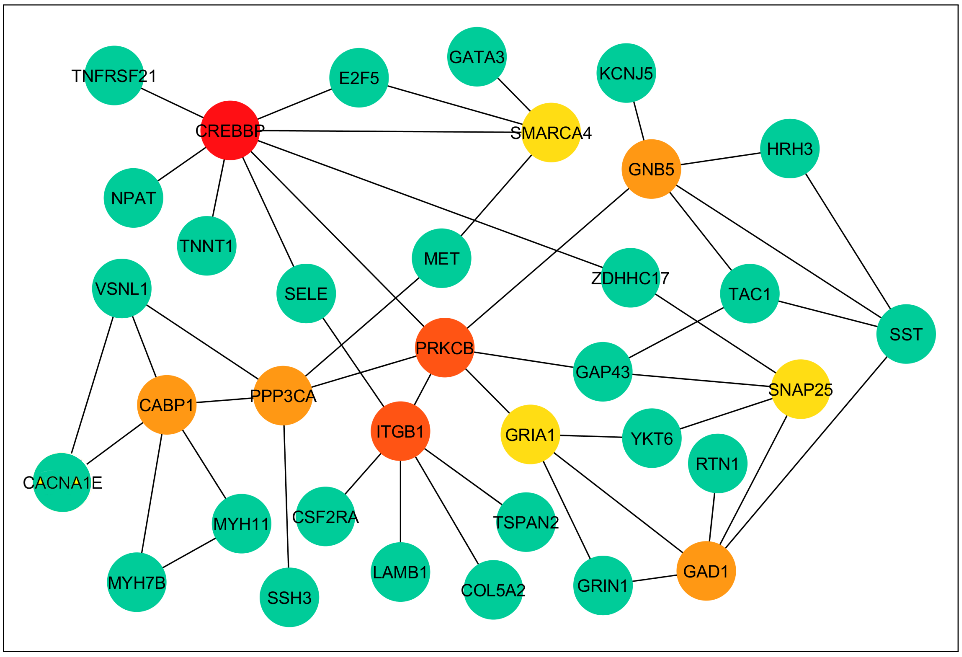

Figure 4.4: The simplified PPI network depicting the hub genes. The 10 most significant hub genes are marked as red, orange and yellow respectively.

Table 4.1: The 10 most significant hub genes. (Ag = Ageing, T2D = Type II Diabetes, Ob = Obesity, Sm = Smoking, HFD = High Fat Diet, RM = Red Meat, AC = high alcohol Consumption, SL = Sedentary Lifestyle.)

Protein	**Description**	**Associated risk factor**
CREBBP	CREB Binding Protein	Up-regulated for HFD
PRKCB	Protein kinase C beta type	Down-regulated for T2D, AC
ITGB1	Integrin beta-1	Down-regulated for HFD
GAD1	Glutamate decarboxylase 1	Down-regulatedfor HFD, Ag
GNB5	Guanine nucleotide-binding protein subunit beta-5	Down-regulated for RM
PPP3CA	Serine/threonine-protein phosphatase 2B catalytic subunit alpha isoform	Down-regulated for Ob
CABP1	Calcium-binding protein 1	Down-regulated for HFD, AC
SMARCA4	Transcription activator BRG1	Up-regulated for T2D
SNAP25	Synaptosomal-associated protein 25	Down-regulated for Ag
GRIA1	Glutamate receptor 1	Down-regulated for Ag

autonomic nervous system development (GO:0048483), sympathetic nervous system development (GO:0048485), neuromuscular process controlling balance (GO:0050885), neuron apoptotic process (GO:0051402), regulation of neurotransmitter transport (GO:0051588) and neuroepithelial cell differentiation (GO:0060563) (see Table 4.3).

4.2.4 Discussion

In this study, we sought to identify novel molecular mechanisms that may affect AD and could be made evident by these gene expression associations with that of known AD risk factors. For this purpose, we conducted analysis in gene expression of AD patients, molecular key pathways, gene ontologies and PPIs. These analyzes that em-

Table 4.2: Some significant KEGG pathways that are related to the nervous system and common among the AD and other risk factors and diseases. (Ag=Ageing, T2D=Type II Diabetes, Ob=Obesity, Sm=Smoking, HFD=High Fat Diet, RM=Red Meat, AC=Alcohol Consumption, SL=Sedentary Lifestyle.)

KEGG ID	Pathway	Genes in pathway	Risk fac./dis.	Adj.P-val
hsa04720	Long-term potentiation	GRIA1, PRKCB, GRIN1, CREBBP, PPP3CA	Ag, Ob, T2D, HFD	5.94E-03
hsa05014	Amyotrophic lateral sclerosis (ALS)	GRIA1, NEFL, NEFM, NEFH, PPP3CA	Ag, AC, Ob, SL	2.80E-04
hsa04728	Dopaminergic synapse	KCNJ5, COMT, GNAL, PRKCB, GNB5	AC, T2D, RM	1.39E-02
hsa05031	Amphetamine addiction	GRIA1, PRKCB, GRIN1, PPP3CA	Ag, T2D, Ob	6.12E-03
hsa04662	B cell receptor signaling pathway	PRKCB, CD22, PPP3CA	T2D, Ob	7.23E-03
hsa04940	Type I T2D mellitus	GAD1, PTPRN2	Ag, HFD, SL	2.76E-02
hsa05100	Bacterial invasion of epithelial cells	ITGB1, MET	HFD, Sm	2.69E-02
hsa05140	Leishmaniasis	HLA-DRB4, PRKCB, ITGB1	T2D, HFD	2.70E-04
hsa05146	Amoebiasis	GNAL, PRKCB, LAMB1,	T2D, Ob	1.32E-02
hsa00250	Alanine, aspartate and glutamate metabolism	FOLH1, GAD1	Ag, HFD	3.00E-04
hsa04014	Ras signaling pathway	PRKCB, RASAL2, GRIN1, GNB5	T2D, RM	7.22E-03
hsa04310	Wnt signaling pathway	PRKCB, PPP3CA, WIF1	Ob	2.17E-03
hsa04360	Axon guidance	ITGB1, MET	Sm	4.96E-02
hsa04370	VEGF signaling pathway	PRKCB, PPP3CA	Ob	3.01E-02
hsa04512	ECM-receptor interaction	ITGB1, LAMB1	Ob	4.02E-02
hsa04514	Cell adhesion molecules (CAMs)	HLA-DRB4, SELE, CD22, ITGB1	T2D	1.94E-03
hsa04721	Synaptic vesicle cycle	SNAP25, SNAP25	Ag, Sm	2.49E-02
hsa04723	Retrograde endocannabinoid signaling	PRKCB, GNB5	RM	2.01E-02
hsa04727	GABAergic synapse	PRKCB, GNB5	RM	1.74E-02
hsa04730	Long-term depression	GRIA1, PRKCB	Ag	2.08E-02

ploy network-based approach can uncover novel relationships between AD and other susceptibility/risk factors. The findings presented here have not been identified by any previous studies. We identified several significant genes that may be usefully investigated in other further work, and the hub genes may identify targets for therapeutic interventions in AD. Besides this, our analysis also identified and characterized a number of biological functions related to these genes that throw light on processes that lead to AD.

Our gene expression analysis showed that there are gene expression patterns that are in common to AD and T2D (35 genes), a total of 34 genes that are dysregulated in both AD and alcohol consumption, these that lie on pathways that may indicate an interaction between these two diseases. Datasets from sedentary lifestyle (18 genes)

Table 4.3: Significant GO ontologies that are related to nervous system, and common between the AD and other risk factors and diseases. (Ag=Ageing, T2D=Type II Diabetes, Ob=Obesity, Sm=Smoking, HFD=High Fat Diet, RM=Red Meat, AC=Alcohol Consumption, SL=Sedentary Lifestyle.)

GO ID	**Pathway**	**Genes in pathway**	**Risk fac./ dis.**	**Adj.P-val**
GO:0045110	Intermediate filament bundle assembly	NEFL, NEFM, NEFH	Ag, AC, Ob, SL	3.95E-05
GO:0002455; GO:0006909; GO:0006911; GO:0006958; GO:0050851; GO:0050853; GO:0050864; GO:0050871; GO:0051251	Humoral immune response mediated by circulating immunoglobulin; Phagocytosis; Phagocytosis, engulfment; Complement activation, classical pathway; Antigen receptor-mediated signaling pathway; B cell receptor signaling pathway; Regulation of B cell activation; Positive regulation of B cell activation; Positive regulation of lymphocyte activation	IGHG3, IGHM, IGHG1, IGHV4-31, IGHD, IGHA1, IGHA2	AC, T2D, SL	5.13E-11
GO:0006836	Neurotransmitter transport	SNAP25	Ag, Sm	6.78E-03
GO:0050890	Cognition	CHRNA4, HRH3	SL, Sm	1.63E-02
GO:0050885; GO:0060563	Neuromuscular process controlling balance; Neuroepithelial cell differentiation	USH1C	Ag	8.22E-03
GO:0046928; GO:0048168; GO:0048935; GO:0051588	Regulation of neurotransmitter secretion; Regulation of neuronal synaptic plasticity; Peripheral nervous system neuron development; Regulation of neurotransmitter transport	MCTP1	T2D	1.56E-02
GO:0007274	Neuromuscular synaptic transmission	CHRNA4	SL	1.86E-02
GO:0007422	Peripheral nervous system development	NFASC	Sm	9.16E-03

and ageing (15 genes) similarly showed evidence of linkage with AD through dysregulated genes. We constructed and analyzed PPI networks to have a better understanding of the mechanism that may be important to either causing or worsening AD. We constructed a PPI network around the DEGs for our study, combining the results of statistical analyzes with the protein interactome network. For finding central proteins (i.e., hubs), topological analysis strategies were employed. These identified Hubs proteins may indicate candidate biomarkers for AD, although these would be of little clinical value (since tissue samples from brain are only available post mortem), they may help to segregate AD classification into subtypes that may be useful. Alternatively, if blockade of the proteins produced from these genes is may feasibly interfere with pathogenic processing these may constitute potential drug targets. From the DEG analysis, 3 up-regulated genes (HLA-DRB4, IGH and IGHA2) and 2 down-regulated genes (IGHD and IGHG1) were common among T2D, Alcoholism and AD. HLA-DRB4 (major histocompatibility complex, class II, DR beta 4) gene has been identified to be associated with AD progression in Genome-wide association studies and meta-analyses [116]. IGH (immunoglobulin heavy locus) is an oxidative modified protein suggesting that the carbonylated level of this gene as associated with AD [117]. From the PPI network analysis, it is observed that 10 hub genes (CREBBP, PRKCB, ITGB1, GAD1, GNB5, PPP3CA, CABP1, SMARCA4, SNAP25 and GRIA1) are involved in the AD. These genes may be dysregulated due to the actions of the risk factors if they are present

Table 4.4: Gene-disease association analysis of DEGs of 7 risk factors and type II diabetes with AD using OMIM and dbGaP databases.

Risk factor/ disease	**Genes**	**Adj.P-val**
Ageing	PLAG1, HMGA2, DNAH11, CR1, VSNL1, DCHS2, F13A1, DISC1, SLC28A1	4.34E-02
Alcohol Consumption	GNAQ, GNAS, ARNT, APBB2, ADCY2, ADCY1, IGF1R, MS4A6A, RTN1, ATXN1, PIEZO2, ST3GAL1	4.72E-01
Type II Diabetes	AKAP13, HNF4A, HMGA2, BUB1, IGF1R, CR1, DIAPH3, TENM4, CADPS, NEDD9, NPAS3	4.66E-01
HFD	AKAP13, CREBBP, HNF4A, HMGA2, DNAH11, COL22A1, PIEZO2, RORA, GFRA2, CD33	5.17E-01
Obesity	RYR2, RBFOX1, VSNL1, NR2F1, HMGA2	1.94E-01
Red Meat	HFE, HMGA2, ADCY2, F13A1	4.41E-01
Sedentary Lifestyle	TFAP2A, HSP90AA1, CREB1, THRA, HFE, APOE, TSHR, RYR2, DIAPH3, DCHS2, DBT, CLU	1.01E-01
Smoking	A2M, OPRD1, SMAD1, ACE, BUB1, CNTNAP2	8.45E-02

(Table-4.1), since we have documented association of these factors with AD, but they may simply be part of pathogenic pathways for AD; if these are indeed the conduits for influences of the risk factors this would indicate that they are important pathways in AD. It was notable that the most significantly identified hub protein CREBBP (CREB binding protein) plays a major role during the evolution of the central nervous system and alteration of CREBBP activity already clearly implicated in AD progression [118]. Its identification here suggests at least utility of our approach.

We have also analyzed our AD-risk factor genes of interest with OMIM and dbGaP databases using EnrichR to validate them by examining other evidence for their involvement in the gene-disease associations. Table-4.4 shows the evidence for these genes according to the current state of knowledge. These results indicate 8 of the AD-risk factor genes of interest have known pathogenic involvement.

The hub genes that we have identified mostly have links to brain physiology and various forms of dementia. CREB-binding protein (CREBBP) is widely expressed and involved in cAMP signaling in many cell types. Point mutations in this gene involved in Rubinstein-Taybi syndrome which affects stature, cancer incidence and learning difficulties associated [119]. CREBBP has been linked to AD in genetic studies [120]. We also previously identified this as a hub gene in AD [121], although that study did not use any datasets used in the present work. Why such a common signalling pathway molecule should be important in this disease is unclear, and is a question around several other hub proteins we identified. However, CREBBP binding partner CREB has shown associations with neurological problems and affects production of amyloid-beta [122] which maybe pathogenic in AD. Similarly, PRKCB, which encodes protein kinase C beta is widely distributed in expression. However, mutations in the gene have no known association with AD although than one study that the gene is highly expressed in blood cell of AD patients so may be a disease marker [123]. The gene is more commonly associated with cancer incidence than neuropathy, although the closely related gene and signalling protein PRKCA is involved in AD development [124].

Hub gene ITGB1 is an integrin cell adhesion protein subunit, integrin beta 1. It is widely expressed, notably in leukocytes and neurones, where it mediates cell-matrix interactions, and binds to laminins in brain tissue in a similar way that (AD associated) amyloid proteins do [125] suggesting that the amyloid may interfere with important roles of ITGB1 in neuronal guidance. ITGB1 is also highly expressed in the hippocampus and blood cells of AD patients [126] so it may relate to the role of inflammation in this disease. GAD1 encodes a glutamate decarboxylase enzyme, but methylation of the gene [127] in brain tissue changes over lifespan and with AD.

GNB5 is a Guanine nucleotide-binding protein subunit involved in guanine activated protein signaling, including dopamine responses seen in neurons, and mutations in the gene are associated with a neuropsychiatric disorder that affects cognition [128]. PPP3CA is a catalytic subunit of calcineurin (Serine/threonine-protein phosphatase 2B) involved in NFATc1 and calcium signal cell pathways and is a target of blocking drugs that suppress the immune system. Splice variants of PPP3CA are associated with AD onset [129] and the protein has wider functions and association in dementia [130]. Hub gene CABP1 is a calcium binding protein with no known link to AD but note that its calcium binding function may intersect with the calcineurin pathways above.

SMARCA4 is a bromodomain protein transcription activator involved in regulating the epigenome and port-translational modifications [131], but not known for a role in brain physiology other than in gliomas [132]. SNAP25 is a synaptosome related protein [133] involved with many types of neuropathology, including AD. Indeed, levels of this protein (and others that form protein complexes with it in the neuronal synapse) can predict cognitive decline in AD patients [134]. Lastly, the GRIA1 gene encodes an important glutamate receptor (part of the glutaminergic neurotransmitter pathway) in the central nervous system with links to migraine and schizophrenia [135] and its levels are altered in brain tissues of AD [136].

Thus in summary, the pathway hub proteins that we have identified in this work have mostly been shown to play an important role in brain physiology and dementia; several are being investigated as drug targets. This supports our network approach to identify genes with an important role in our disease of interest.

4.3 Experiment Name: Systems biology and bioinformatics approach to identify gene signatures, pathways and therapeutic targets of Alzheimer's disease

4.3.1 Short Summary

Alzheimer's disease (AD) develops relentlessly in affected individuals and its occurrence is increasing. A clinical test to diagnose early-stage AD could be an important means of enabling interventions to slow its progression. However, available neuroimaging and cerebrospinal fluid-based diagnoses are very costly. Therefore, detecting AD from blood transcripts that mirror the expression of brain transcripts in the AD could improve the diagnosis. To achieve this goal, we employed a transcriptional analysis of affected tissues and integrated them with cis-eQTL data. In this study, we analyzed microarray gene expression data of brain and blood cells from AD patients and control individuals. Differentially expressed genes (DEGs) common to both brain tissue and blood cells were identified. Potential common genes and molecular pathways were identified using overlapping DEGs through the pathway and gene ontology enrichment analysis. We identified 18 significantly dysregulated genes shared by both brain and blood cells in AD affected individuals. We validated these candidates as disease-associated genes using gold-standard benchmarking databases (gene SNP-disease linkage). Significant molecular pathway and gene ontology indicating AD progression were identified. This study also identified regulatory factors, including transcription factors (TFs), microRNAs and candidate drugs. In sum, we identified new putative links between pathological processes in brain tissue and blood cells in AD that may allow assessment of AD status using blood samples. Thus, our formulated methodologies demonstrate the power of gene and gene expression analysis for brain-related pathologies transcriptomics, cis-eQTL, and epigenetics data from brain and blood cells.

4.3.2 Materials and Methods

Dataset

We collected two different gene expression microarray datasets of AD from the Gene Expression Omnibus of the National Center for Biotechnology Information (NCBI) (`http://www.ncbi.nlm.gov/geo/`) with accession numbers GSE18309 (blood) and GSE1297 (brain) for our study. GSE18309 is an Affymetrix human genome array data of peripheral blood mononuclear cell (PBMC) transcriptomes from 4 AD patients and 4 control individuals. The AD dataset (GSE1297) is obtained by gene expression profiling of hippocampal tissues on 31 separate microarrays from nine control subjects and 22 AD patients with varying severity [105]. We also studied the eQTL data for both brain and blood from GTEx Portal database (`https://gtexportal.org/home/`).

Analytical Approach

In this study, we have employed a systemic and quantitative approach to detect AD at an early stage by identifying the brain dysregulation in blood. Fig. 4.5 illustrates the procedure whereby gene expression microarray data, Genome-Wide Association Studies (GWAS), and expression of quantitative trait loci (eQTL) datasets are used. In order to identify the significant components of molecular pathways for AD that are common between the brain and blood, we have performed the gene over-representation analysis, Gene Ontology (GO) and signaling pathway enrichment analysis, disease-gene correlations and protein-protein interaction (PPI). This study also employs biomarker signatures at both transcriptional (mRNAs and miRNAs) and translational levels (hub proteins and TFs) and candidate drug identification. The gold benchmark validated datasets dbGaP (`https://www.ncbi.nlm.nih.gov/gap/`), DisGeNET [137], GWAS_Catalog_2019 (`https://www.ebi.ac.uk/gwas/home`) and UK_Biobank_GWAS_v1 (`https://www.ukbiobank.ac.uk/tag/gwas/`) were included in our study to validate the principle of our approach.

4.3.3 Results

DEGs analysis of datasets

The DEGs analysis of the brain and blood datasets identified 484 (336 up and 148 down) and 1702 (736 up and 966 down) DEGs, respectively. The overlapping DEGs

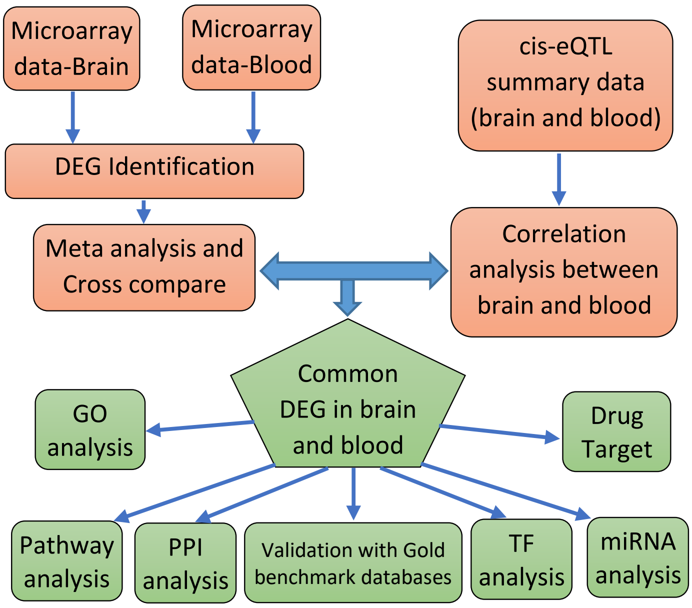

Figure 4.5: A pipeline to identify blood cell-expressed Alzheimer's disease biomarkers and pathways that are coordinately expressed in brain tissue.

between brain and blood obtained through the cross comparative analysis are graphically depicted in Fig. 4.6 and Fig. 4.7. 20 common up-regulated DEGs and 9 common down-regulated DEGs along with their log fold change (log_2FC) and negative logarithmic FDR adjusted p-values are shown in Fig. 4.6 and Fig. 4.7, respectively. The expression levels of these 29 common DEGs in blood and brain tissues from AD patients and normal individuals are provided in the supplementary tables (Table S1 and S2).

Identifying genes expressed in blood cells that mirror those expressed in brain

We identified genes having similar genetic control of expression in blood and brain cells incorporating a correlation and meta-analysis approach as explained in the method section. Very few eQTL databases are available linking gene SNPs to gene expression, among them we used the GTEx database in this study. Our correlation and meta-analysis approach identified 673 blood-brain co-expressed genes (BBCG).

Identifying genes in the blood that influence AD development

Utilizing the curated gold-benchmark OMIM database and GWAS catalogues, we identified the AD-associated genes using their expression patterns (e.g. SNPs). We identified 18 significant genes for AD which were commonly dysregulated between blood and brain. These are HSD17B1, GAS5, RPS5, VKORC1, GLE1, WDR1, RPL12, MORN1, RAD52, SDR39U1, NPHP4, MT1E, SORD, LINC00638, MCM3AP-AS1, GSDMD, RPS9, and GNL2. However, these genes exhibited no overlap with the common DEGs in blood and brain indicating a limitation of this study. We further identified regulating pathways using these potential AD biomarkers.

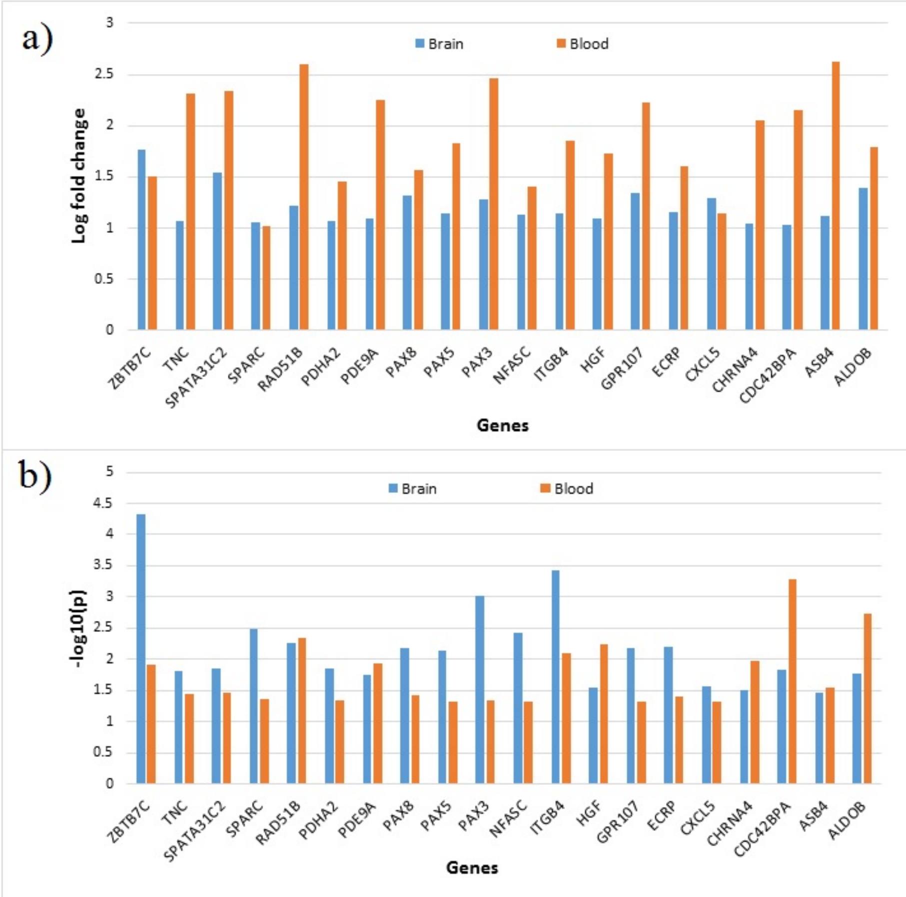

Figure 4.6: Identified upregulated DEGs in both brain and blood. A) log fold changes of the common upregulated significant genes in brain and blood and B) Negative logarithmic FDR adjusted p-values of upregulated significant genes common to brain and blood.

Identifying pathways in blood cells that mirror in brain

We performed pathway analysis on the common DEGs between blood and brain using the KEGG pathway database as well as the web-based gene set enrichment analysis tool EnrichR. A variety of GO terms and pathways were enriched, including 122 biological processes (BP), 28 molecular functions (MF), 25 cellular components (CC) and 10 KEGG pathways. The top 5 GO terms of BP, MF, CC and significant KEGG pathways are shown in Table 4.5 and Table 4.6 respectively.

Protein-protein interaction (PPI) analysis to identify functional sub-networks

We constructed the PPI network using the web-based visualization resource STRING [138] with a medium confidence level (0.4) based on the common disease gene sets. Markov clustering algorithm was incorporated to cluster the genes in the network. Malfunction of a protein complex may cause several diseases that can be identified by the dysfunction of the protein subnetwork. Again, the association of several proteins in the PPI network reveals the potential association of several genes. Hence, we looked for sub-networks in the PPI network using the genes engaged in the common pathway and processes of blood and brain. The resulted PPI shown in Fig. 4.8 confirms the existence of PPI sub-network in our enriched geneset and hence endorse the availability

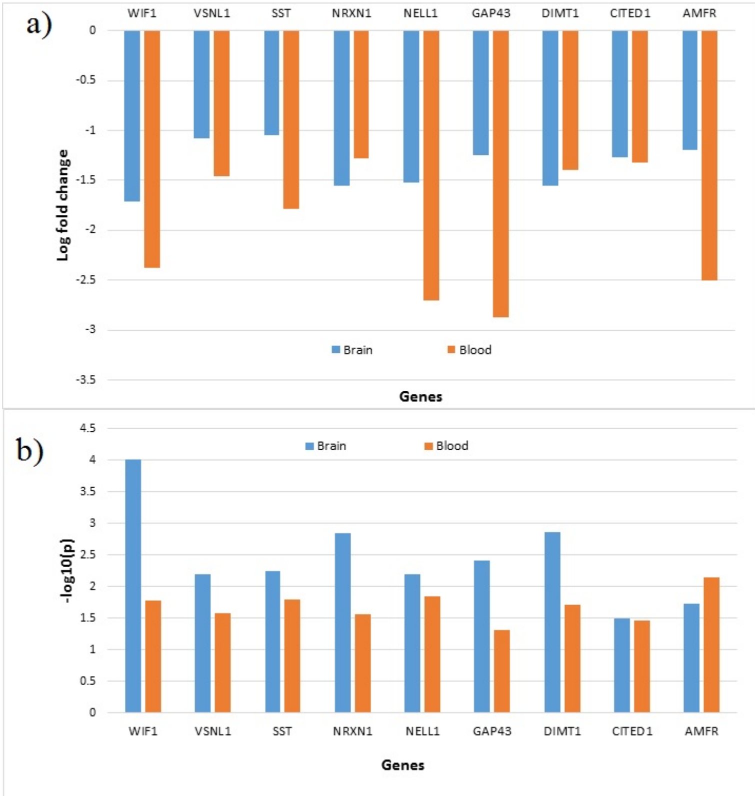

Figure 4.7: Identified downregulated DEGs in both brain and blood. A) log fold changes of the common downregulated significant genes in brain and blood and B) Negative logarithmic FDR adjusted p-values of downregulated significant genes common to brain and blood.

of related functional pathways. A simplified PPI network was constructed using topological analysis by Cyto-Hubba plugin to show the 10 most significant hub proteins (Fig. 4.9), which are SST, GAP43, NRXN1, CHRNA4, VSNL1, HGF, WIF1, PAX3, NFASC and DIMT1.

Identification of Post-transcriptional Regulator

We identified TFs and miRNAs and their targeted DEGs to reveal regulatory biomolecules that may regulate the expression of DEGs at transcriptional and post-transcriptional levels (Table 4.7). The analysis revealed significant TFs (RORB, NR2F1, HNF1B, NKX2-8, NFAT2 and MZF1) and miRNAs (Table 4.8) played significant roles in the regulation of the DEGs identified this study.

Drug target and candidate drug identification

After manual curation with criteria of adjusted p-value less than 0.05 we identified 29 significant drug targets using the DSigDB database and the top 10 drug target and candidate drugs are shown in Table 4.9.

Table 4.5: The top 5 significant GO terms for common DEGs between blood and brain tissue in AD.

Category	GO Terms	Pathway	Adj.P-val	Genes
BP	GO:0042254	ribosome biogenesis	4.00E-05	RPS9, RPL12, RPS5, GNL2
	GO:0006614	SRP-dependent cotranslational protein targeting to membrane	7.00E-05	RPS9, RPL12, RPS5
	GO:0006613	cotranslational protein targeting to membrane	8.00E-05	RPS9, RPL12, RPS5
	GO:0045047	protein targeting to ER	9.00E-05	RPS9, RPL12, RPS5
	GO:0007399	nervous system development	2.06E-02	NFASC, NRXN1
MF	GO:0019843	rRNA binding	8.00E-06	RPS9, RPL12, RPS5
	GO:0004303	estradiol 17-beta-dehydrogenase activity	7.17E-03	HSD17B1
	GO:1901612	cardiolipin binding	7.17E-03	GSDMD
	GO:1901611	phosphatidylglycerol binding	9.85E-03	GSDMD
	GO:0070300	phosphatidic acid binding	1.52E-02	GSDMD
CC	GO:0005840	ribosome	4.00E-05	RPS9, RPL12, RPS5
	GO:0022626	cytosolic ribosome	1.80E-04	RPS9, RPL12, RPS5
	GO:0044445	cytosolic part	3.70E-04	RPS9, RPL12, RPS5
	GO:0022627	cytosolic small ribosomal subunit	9.10E-04	RPS9, RPS5
	GO:0015935	small ribosomal subunit	1.06E-03	RPS9, RPS5

Table 4.6: Significant KEGG pathways for common DEGs between blood and brain tissue in AD.

Terms	Pathway	Adj.P-val	Genes
hsa03010	Ribosome	2.30E-04	RPS9, RPL12, RPS5
hsa00130	Ubiquinone and other terpenoid-quinone biosynthesis	9.85E-03	VKORC1
hsa03440	Homologous recombination	2.57E-02	RAD52
hsa00051	Fructose and mannose metabolism	2.84E-02	SORD
hsa00040	Pentose and glucuronate interconversions	3.19E-02	SORD
hsa04913	Ovarian steroidogenesis	4.41E-02	HSD17B1
hsa04978	Mineral absorption	4.49E-02	MT1E

Table 4.7: Significant transcription factors regulating common DEGs between blood and brain tissue in AD.

Terms	Genes	Adj.P-val
RORB	GLE1, GSDMD, CITED1, VSNL1, AMFR, TNC, PAX5	1.00E-02
NR2F1	GLE1, GSDMD, CITED1, VSNL1, AMFR, TNC, PAX5	1.21E-02
HNF1B	GAP43, SPARC	1.89E-02
NKX2-8	PAX8, VSNL1, ITGB4	2.22E-02
NFAT2	NELL1, PDHA2, GAP43, VSNL1, HGF, SDR39U1, ALDOB	4.04E-02
MZF1	GSDMD, CHRNA4, SST, RPL12, NPHP4, ASB4, PDE9A	4.21E-02

Table 4.8: Significant MicroRNAs regulating common DEGs between blood and brain tissue in AD.

Terms	Genes	Adj.P-val
I-miR-210-3p	RAD52, DIMT1, HSD17B1	2.95E-03
I-miR-1237-3p	SORD, GPR107, CXCL5	4.70E-03
I-miR-345-5p	PAX8, SORD	6.40E-03
I-miR-1248	RPS9, SORD, CXCL5	6.63E-03
I-miR-181b-3p	NFASC, ASB4	1.08E-02
I-miR-4648	RPL12, AMFR	1.37E-02
I-miR-4654	RPL12, AMFR	1.44E-02
I-miR-4769-5p	RPL12, AMFR	1.52E-02
I-miR-5087	VSNL1, CXCL5	1.63E-02
I-miR-3166	VSNL1, RPS5	1.66E-02
I-miR-484	VKORC1, RAD51B, RPS9, DIMT1, SORD, GNL2	1.73E-02
I-miR-548m	PAX5, CXCL5	2.10E-02
I-miR-518e-3p	VSNL1	2.32E-02
I-miR-8068	PAX8, VSNL1	2.39E-02
I-miR-4743-3p	TNC, GPR107	2.66E-02
I-miR-4756-5p	GPR107, PAX5, CXCL5	2.95E-02
I-miR-1321	GPR107, PAX5, CXCL5	2.98E-02
I-miR-4739	GPR107, PAX5, CXCL5	3.06E-02
I-miR-4496	RAD51B, NFASC	3.18E-02
I-miR-1296-5p	NFASC, RPS9	3.23E-02
I-miR-518c-3p	HSD17B1	3.24E-02
I-miR-302a-5p	NRXN1, SORD	3.54E-02
I-miR-3653-3p	SPARC, PAX5	3.74E-02
I-miR-216b-5p	SORD, CDC42BPA	4.12E-02
I-miR-578	GPR107, CXCL5	4.84E-02
I-miR-200a-3p	RPL12, HGF	4.90E-02

Table 4.9: Top 10 significant drug targets identified for common DEGs between blood and brain tissue in AD.

Drug/Small molecule	Adj.p-val	Genes
Decitabine	2.00E-05	SPARC, PAX8, WIF1, HSD17B1, PAX3
meclofenoxate	2.80E-04	PAX8, ITGB4, NRXN1, SPATA31C2, PAX3, ALDOB
AC1NRCGS	4.80E-04	HSD17B1, SORD
Dibenz[a,h] anthracene	8.50E-04	RAD52, ALDOB, MT1E
TERT-BUTYL HYDROPEROXIDE	9.40E-04	SPARC, WDR1, VSNL1, ITGB4, DIMT1, SST, TNC, SORD, NPHP4, GPR107
Imatinib mesylate	9.60E-04	RAD52, HSD17B1, SORD, PAX3
coumarin	1.00E-03	VKORC1, HSD17B1
zalcitabine	1.09E-03	GAP43, VSNL1, PAX8, CHRNA4, NRXN1, HGF, SPATA31C2, ALDOB
1-METHYLPH ENANTHRENE	1.21E-03	ALDOB, MT1E
Quinomycin A	1.21E-03	RAD52, PAX8

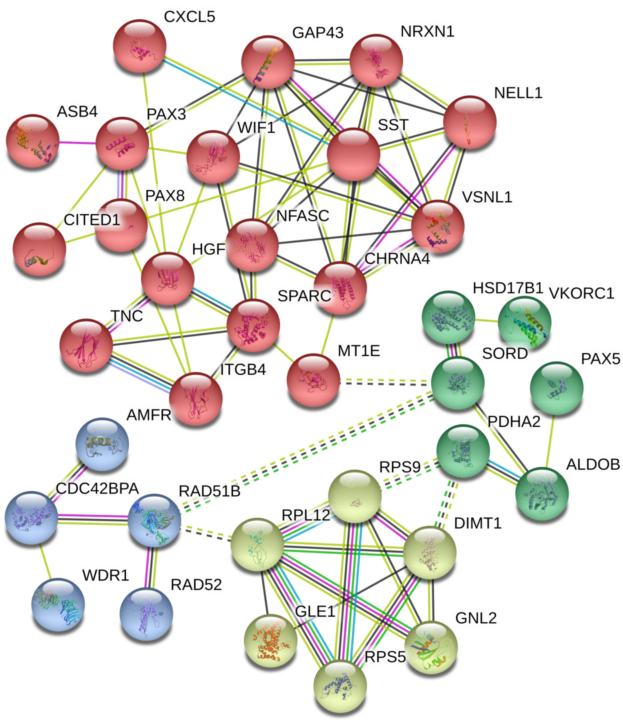

Figure 4.8: Protein-protein interaction (PPI) network of the AD. The nodes indicate the proteins and the edges indicate the interactions between two proteins. Color indicates MCL analysis clusters of proteins.

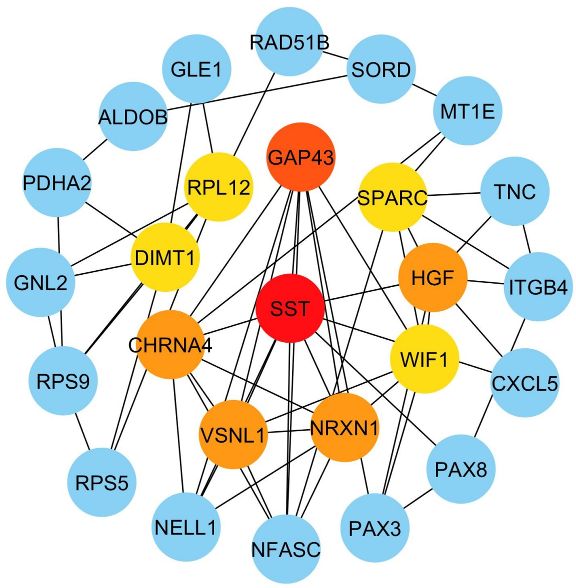

Figure 4.9: The simplified PPI network of the common DEGs between blood and brain AD. The 10 most significant hub proteins are marked as red, orange and yellow respectively.

Table 4.10: Gene-disease association analysis of the potential biomarkers in blood cells for AD using benchmark databases.

Database	**Disorders**	**Adj.P-val**	**Genes**
dbGaP	Alzheimer Disease	2.04E-01	VSNL1
DisGeNET	Alzheimer Disease, Early Onset	3.04E-02	VSNL1, AMFR
	Alzheimer Disease, Late Onset	2.15E-02	VSNL1, NRXN1, AMFR
	Alzheimer's Disease	1.79E-01	GAP43, VSNL1, CHRNA4, SST, HSD17B1, AMFR, HGF
	Alzheimer's Disease, Focal Onset	1.70E-02	VSNL1, AMFR
GWAS_Catalog_2019	Alzheimer's disease in APOE e4+ carriers	4.82E-02	SORD
UK_Biobank_GWAS_v1	Alzheimer's disease	1.40E-02	NPHP4

4.3.4 Discussion

In the present study, we investigated brain and blood transcriptomics and eQTL data to identify interesting common pathways in those cells. We first employed global transcriptomic analysis to determine overlapping DEGs in brain tissue and blood cells as well as potential common pathways. Then we compared these identified pathways with validated datasets dbGaP, DisGeNET, GWAS_Catalog_2019 and UK_Biobank_GWAS_V1 shown in Table 4.10. Finally, the network-based approach using the PPI data demonstrated significant pathways common in the brain and blood.

Among the hub genes revealed in this study, Somatostatin (SST) is directly associated with Aβ peptide in the brain which is the primary influencing AD pathogenesis [139, 140]. The growth-associated protein 43 (GAP-43) is related to both diagnosis and neuropathology of AD and thus can be considered as a potential biomarker for AD [141]. GAP-43 gene causes significant molecular damage that precedes and progresses AD [142]. NRXN1, a genetic variant of the neurexin gene family has an association with autism spectrum disorder (ASD) [143] and other variants of this family plays vital role in AD progression [144]. Single-nucleotide polymorphisms (SNPs) within visinin-like 1 (VSNL1) are evidenced to have strong interconnection with AD cases with psychosis [145]. Hepatocyte growth factor (HGF) exists at a higher level in neurons as well as other tissues of the nervous system and in the cerebrospinal fluid of AD cases [146]. Gradual decrease in the expression of WNT inhibitory factor-1 (WIF1) from healthy people to AD patients results in enhancement of Wnt signaling, reduction of GSK3β function and Tau phosphorylation [147]. Mutations in paired box-3 (PAX3) usually causes Waardenburg syndrome [148] and are also associated with AD progression [149]. Neurofascin (NFASC) is rarely found to be associated with AD but its involvement in synapse formation, plasticity, and stability is evidenced [150].

We also identified TFs and miRNAs having a significant influence on gene expression at the transcriptional and post-transcriptional level, that implies their role as potential biomarkers. RAR related orphan receptor B (RORB) is recently manifested as responsible for the selective vulnerability of neurons in the entorhinal cortex that characterizes AD [151]. Nuclear receptor subfamily 2 group F member 1 (NR2F1) dysregulation is responsible for optic atrophy leading intellectual ailment [152]. HNF1 homeobox B (HNF1B) corresponds to type II diabetes especially in elder people [153]. SNPs in myeloid zinc finger 1 (MZF1) regulating inflammatory response and cholesterol metabolism employ association with AD [154]. miR-210-3p is reported as a putative biomarker for AD in [155]. Up-regulation of miR-345-5p is evidenced with AD [156] whereas miR-345-5p is associated with both major depression and bipolar disorder [157]. The reported candidate drug component meclofenoxate has been used to enhance memory or other cognitive functions in senile dementia and AD [158]. In [159], a series of novel tacrine-coumarin hybrids were reported as multi-target agents against AD.

4.4 Experiment Name: System biology and bioinformatics pipeline to identify comorbidities risk association: neurodegenerative disorder case study

4.4.1 Short summary

A wide spectrum of comorbidities, including other neurodegenerative diseases (NDDs), are frequently associated with AD. How AD interacts with those comorbidities can be examined by analysing gene expression patterns in affected tissues using bioinformatics tools. We surveyed public data repositories for available gene expression data on tissue from AD subjects and from people affected by neurodegenerative diseases that are often found as comorbidities with AD. We then utilized large set of gene expression data, cell-related data and other public resources through an analytical process to identify functional disease links. This process incorporated gene set enrichment analysis and utilized semantic similarity to give proximity measures. We identified genes with abnormal expressions that were common to AD and its comorbidities, as well as shared gene ontology terms and molecular pathways. Our methodological pipeline was implemented in the R platform as an open-source package and available at the following link: `https://github.com/unchowdhury/AD_comorbidity`. The pipeline was thus able to identify factors and pathways that may constitute functional links between AD and these common comorbidities by which they affect each others development and progression. This pipeline can also be useful to identify key pathological factors and therapeutic targets for other diseases and disease interactions.

4.4.2 Materials and methods

Dataset

We obtained gene expression datasets from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) and European Bioinformatics Institute Array Express database. We queried for AD and found 531 datasets, most of them were disqualified at the start by being very low sample size compared to our selected cut off sample size 10, duplicate datasets, having inappropriate format or undesirable experimental set-up, RNAseq datasets, and from organisms other than human. Thus we selected 8 datasets to be highly relevant to AD and appropriate for our study. The finally selected gene expression datasets for AD have the accession numbers: GSE1297, GSE110226, GSE33000, GSE48350, GSE12685, GSE5281, GSE4229 and GSE4226. All datasets were generated using central nervous system tissues and Affymetrix array platforms except GSE4226 and GSE4229 which were MGC arrays of peripheral blood analyses. GSE1297 is a correlation analysis of hippocampal tissues from nine control subjects and 22 AD patients with varying severity [105]. GSE110226 compared transcripts of choroid plexus from postmortem tissues of 6 healthy samples and 7 AD patients, 4 FTD patients and 3 HD patients [160]. GSE33000 analysed post mortem prefrontal cortex tissues of 310 AD patients, 157 HD patients and 157 non-demented samples [161]. GSE48350 is the profiling of hippocampus, entorhinal cortex, superior frontal cortex and post-central gyrus regions in 170 healthy individuals and 80 AD cases [162]. GSE12685 is a comparative study of gene expression for frontal cortex synaptoneurosomes between 6 normal controls and 8 AD patients [163]. GSE5281 is obtained by analyzing 16 unaffected and 19 AD affected tissues, specifically 6 central nervous system tissues: entorhinal cortex, hippocampus, medial temporal gyrus, posterior cingulate, superior frontal gyrus and primary visual cortex cells [164]. GSE4229 is a study of genetic variations of peripheral blood mononuclear cells from 22 healthy old people and 18 AD cases using the NIA Human MGC cDNA microarray [165]. GSE4226 compares peripheral blood mononuclear cells obtained from 14 normal elderly control (NEC) and 14 AD affected subjects [166]. For the study of neurodegenerative comorbidity analysis of AD we selected GSE7621, GSE6613, GSE49036 and GSE54536 for PD; GSE93767, GSE110226 and GSE33000 for HD; GSE833 and GSE107375 for ALS; GSE27206 for SMA; GSE49036 for LBD; GSE110226, GSE13162 and GSE40378 for FTD; GSE21942 for MS. GSE7621 is generated by extracting RNA from substantia nigra tissue of postmortem brain of 9 controls and 16 PD patients and hybridizing on Affymetrix microarrays [167]. GSE6613 is whole blood expression data analysis from PD patients and controls [168]. GSE49036 is an overall study of gene expression of subtantia niagra tissue from PD patients, LBD cases and normal individuals [169]. GSE54536 is obtained through a whole-transcriptome comparison of the peripheral blood from PD patients with healthy subjects [170]. GSE93767 is a transcriptional analysis of human-induced pluripotent stem cells (hiPSC) using a CRISPR-Cas9 from HD cases compared with controls [171]. GSE833 is a gene expression profiling of grey matter from post mortem spinal cord of ALS patients and controls [172]. GSE107375 is a whole transcriptome expression analysis of the motor cortex from 10 controls and 30 ALS cases [173]. GSE27206 is the gene expression data evaluation of induced pluripotent stem cells (iPS cells) for SMA [174]. GSE13162 is obtained through global expression profiling using a microarray of postmortem brain cells from the frontal cortex, hippocampus, and cerebellum [175]. GSE40378 is a gene expression analysis by an array of induced pluripotent stem cell models [176]. GSE21942 is a comparison of the expression level of genes for peripheral blood mononuclear cells between MS patients and controls [177].

Analytical Approach

At first, the chosen gene expression datasets and their matrix information were downloaded and converted to Expression Set class for differential gene expression analysis. We reviewed the sample records (GSM) manually for sample classification and constructed design models (patients, controls). The created design model for AD cases is AD patient vs healthy individual and patient of neurodegenerative diseases vs healthy control for other cases. These design models are then filtered using a linear and a Bayesian method. Using a threshold for p-value and absolute log Fold Change (logFC) values to be at most 0.05 and at least 1.0 respectively, DEGs are identified.

We constructed the topGOdata class using the selected genes by specifying the GO domain and stipulating the annotation to perform the mapping. We then obtained the filter for GO terms and their associations with the DEGs by employing the Fisher's exact test. After that, we performed the semantic similarity comparison among all the selected diseases considering DEGs, GO terms and DO terms to measure the proximity for all the chosen datasets. The DO terms used in this study for the corresponding diseases are AD DOID: 10652, PD DOID: 14330, HD DOID: 12858, ALS DOID: 332, SMA DOID: 12377, LBD DOID: 12217, FTD DOID: 9255 and MS DOID: 2377. We then performed the KEGG pathway analysis for the DEGs to find out significant molecular pathways or diseases for AD and its comorbidity datasets. Finally, the statistical information, genes-GO term associations, DAGs, semantic similarity measures along with dendrograms for DEGs, GO terms and DO terms are generated as final output. Furthermore, we generated a gene network using the common DEGs between AD and its comorbidities, with enlightenment on the pathways/diseases. Figure 4.10 pictures the block diagram of the analytical process. Various BioConductor 3.4 R packages [178] were used to develop the analytical approach. We downloaded the selected datasets from the NCBI GEO and converted the data into form Expression Set class using GEOquery 2.40.0. GEOquery offers corresponding methods to access various types of GEO data [179]. Linear Models for Microarray Data (limma) 3.30.8 was used for differential gene expression analysis by comparing the transcriptomic profiles of healthy subjects with that of the patients. Limma provides compact collection of tools to analyze gene expression microarray data [180]. We filtered the genes using genefilter 1.56 for the threshold values p-value less than 0.05 and absolute logFC greater than 1. Genefilter offers necessary methods to curate genes obtained in high throughput experiments [181]. We incorporated the topGO 2.26 for the enrichment analysis for GO and performed the Fisher's exact test to obtain the topology of the DAG [182]. The semantic similarity between the selected pathologies were determined for GO terms and DEGs using GOSemSim 2.0.4 that serves as a quantitative tool for the semantic comparisons [183]. The semantic similarity for DO terms was evaluated by Disease Ontology Semantic and Enrichment analysis (DOSE) 3.0.10 [184]. Finally, the KEGG pathway enrichment analysis was performed using clusterProfiler 3.2.14, which offers statistical analysis and visualization methods for functional profiles of genes [185]. We used the GEO file transfer protocol (ftp) call to download GEO datasets instead of using GEOquery package due to some interaction issues with other used packages.

4.4.3 Results

Statistical Summary and GO Term Trees

The statistics about all the chosen AD studies are mentioned in Table 4.11. The threshold for p-values is 0.05 and for absolute logFC is 1.0 to obtain the number of genes shown in 4th, 5th and 6th columns from left. The numbers shown in brackets for 6th column are obtained using 2.0 as threshold values of logFC. Similarly, Table 4.12 summarizes the statistics for the selected neurodegenerative comorbid pathologies of AD. Table 4.13 shows the synopsis of the selected datasets along with the number of analyzed DEG.

DAG of GO terms is constructed for each selected pathologies. The graphs manifest that all the GO terms are not trivial and hence are hidden. Figure4.11 shows such a

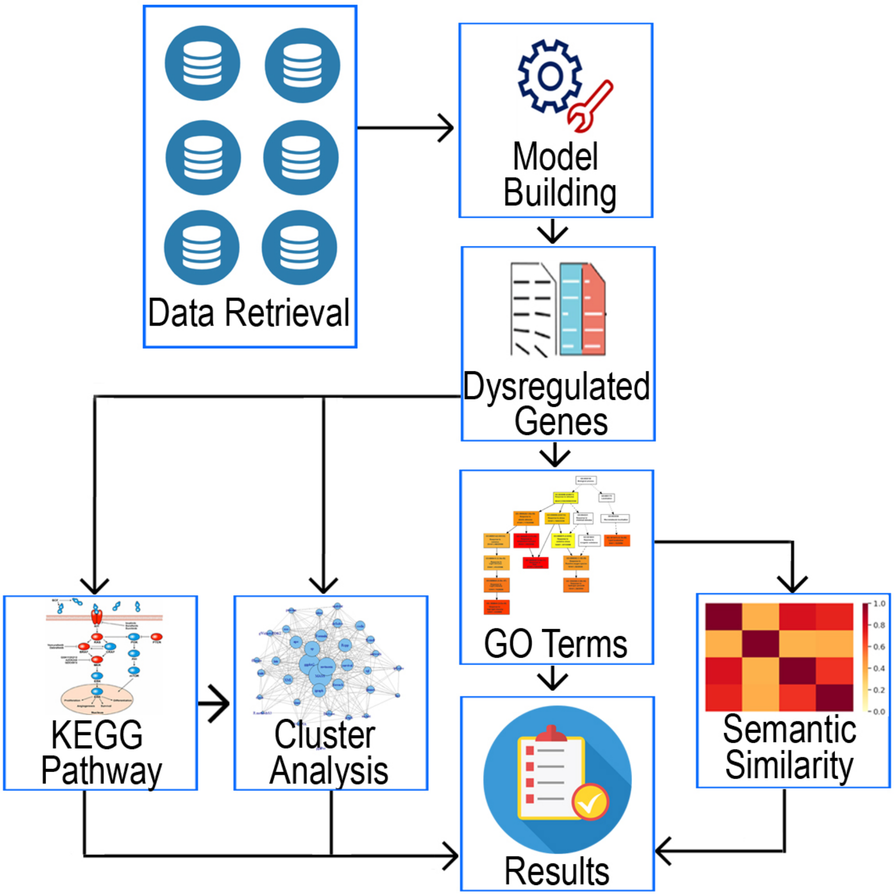

Figure 4.10: Pipeline of the analytical approach.

DAG for the dataset GSE12685 of AD study.

Pathways

The five most significant BP GO terms involved in each AD study are as follows:

i GSE110226: immune system process, regulation of immune system process, positive regulation of immune system process, nitrogen compound metabolic process, and transport.

ii GSE12685: adaptive immune response, antimicrobial humoral immune response, innate immune response, epithelial cell differentiation and extracellular matrix organisation.

iii GSE1297: immune system process, nitrogen compound metabolic process, cell communication, system process, and transport.

iv GSE33000: biological process, nitrogen compound metabolic process, signal transduction, cell communication, and transport.

v GSE4226: reproduction, cell activation, regulation of cell growth, response to active oxygen species and response to the acid chemical.

Table 4.11: Statistical summary for AD studies. The 3rd, 4th, 5th and 6th columns represent the count for unfiltered genes, the significant DEGs with threshold for p-value, adjusted p-value and logFC (numbers in brackets are for logFC with threshold 2). 7th and 8th columns show the count for unfiltered GO terms and filtered GO terms using Fisher test.

Dataset	**Tissue source**	**Genes**	**P-Value**	**Adj. P-Value**	**LogFC**	**GO Terms**	**Fisher test**
GSE110226	Choroid plexus	21003	6002	475	442 (24)	200	11
GSE12685	Frontal cortex synaptoneurosomes	13907	2986	1	180 (0)	211	26
GSE1297	Hippocampal CA1 Tissue	13907	2830	0	565 (10)	156	9
GSE33000	Prefrontal cortex	19518	16105	15858	0 (0)	201	26
GSE4226	Peripheral blood mononuclear	6571	457	0	581 (299)	84	21
GSE4229	Peripheral blood mononuclear	6571	332	0	432 (219)	135	6
GSE48350a	Hippocampus	22832	10222	3515	322 (9)	147	14
GSE48350b	Entorhinal cortex	22832	7002	645	114 (6)	197	7
GSE48350c	Superior frontal cortex	22832	8419	2537	78 (6)	125	6
GSE48350d	Post-central gyrus	22832	5416	435	21 (5)	84	4
GSE5281	Entorhinal cortex, hippocampus, medial temporal gyrus, posterior cingulate, superior frontal gyrus and primary visual cortex	22832	12726	10699	2306 (35)	113	18

vi GSE4229: biological process, metabolic process, nitrogen compound metabolic process, cell communication and signal transduction.

vii GSE48350a: biological process, cellular process, nitrogen compound metabolic process, metabolic process and transport.

viii GSE48350b: nitrogen compound metabolic process, cell communication, system process, response to stress and transport.

ix GSE48350c: biological process, cellular process, metabolic process, regulation of biological process and regulation of the cellular process.

x GSE48350d: cell activation, myeloid leukocyte activation, myeloid cell activation involved in immune response, endothelial cell activation involved in immune response, cell activation involved in immune response and immune effector process.

Table 4.12: Statistical summary for studies of neurodegenerative comorbid diseases of AD. The 4th, 5th, 6th and 7th columns represent the count for unfiltered genes, the significant DEGs with threshold for p-value, adjusted p-value and logFC (numbers in brackets are for logFC with threshold 2). The 8th and 9th columns show the count for unfiltered GO terms and filtered GO terms using Fisher test.

Dataset	Dis.	Tissue source	Genes	P-Value	Adj. P-Value	LogFC	GO Terms	Fisher test
GSE49036	PD	Substantia nigra	22832	6454	67	228 (3)	249	25
GSE6613	PD	Whole blood	13907	1991	0	4 (0)	106	6
GSE7621	PD	Substantia nigra	22787	4389	1	1672 (55)	102	19
GSE54536	PD	Peripheral blood	20760	8466	5855	4009 (1631)	64	22
GSE110226	HD	Choroid plexus	21003	3542	1	313 (12)	76	30
GSE33000	HD	Prefrontal cortex	19518	16328	16144	0 (0)	112	14
GSE93767	HD	Induced pluripotent stem	20053	1245	2	1632 (92)	61	11
GSE49036	LBD	Substantia nigra	22832	3651	0	184 (3)	100	19
GSE68605	ALS	Motor neurons	22832	2596	7	5768 (343)	404	49
GSE833	ALS	Spinal cord	6068	765	19	2555 (931)	343	56
GSE110226	FTD	Choroid plexus	21003	5164	0	629 (29)	77	25
GSE13162	FTD	Frontal cortex, hippocampus, and cerebellum	13907	4771	2099	139 (1)	43	15
GSE40378	FTD	Induced pluripotent stem	20760	3752	565	21 (2)	43	15
GSE21942	MS	Peripheral blood	22832	9379	5876	524 (62)	84	25
GSE27206	SMA	Induced pluripotent stem	22832	2117	0	1225 (232)	99	43

xi GSE5281: nitrogen compound metabolic process, response to stress, cellular aromatic compound metabolic process, nucleobase-containing compound metabolic process and transport.

The DEGs comparison between the AD datasets and its neurodegenerative comorbidities reveals the following overlapping genes: ACTB, CEACAM8, COX2, DEFA4, GFAP, MALAT1, RGS1, RPE65, SYT1, S100A8, S100A9, SERPINA3, TNFRSF11B

Table 4.13: Summary of findings in the steps of the pipeline for the selected pathologies.

Disease	**Tissue source**	**Available dataset**	**Selected dataset**	**Up DEGs**	**Down DEGs**
Alzheimer's Disease	Brain, blood	531	8	2037	1598
Parkinson's Disease	Brain, blood	196	4	961	1345
Huntington's Disease	Brain	64	3	315	418
Lewy Body Disease	Brain	11	1	57	93
Amyotrophic Lateral Sclerosis	Brain, spinal cord	104	2	1563	1666
Frontotemporal Dementia	Brain	28	3	447	278
Multiple Sclerosis	Blood	124	1	213	317
Spinal Muscular Atrophy	Brain	20	1	250	211

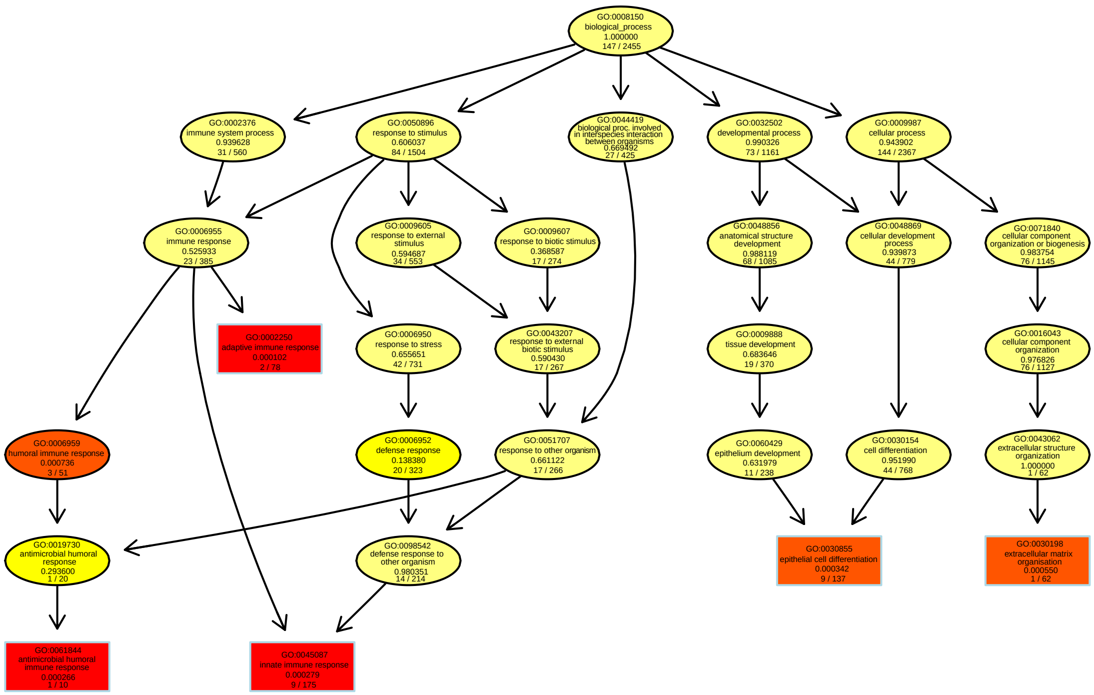

Figure 4.11: Example DAG of GO terms with GSEA on GSE12685 dataset of AD. The original graph (on the top) and a zoom (on the bottom) are presented. The 5 most significantly enriched GO terms are indicated by the rectangles and the oval shaped nodes represent significant GO terms. The red and orange colors indicate the most significant GO terms. The last two lines inside each node show raw p-value followed by the number of significant genes and the total number of genes annotated to the corresponding GO term for the dataset.

and TUBB2A. We built a cluster network for these overlapping DEGs using the online tool GeneMania [186]. For this we took physical interactions, co-expression, predicted, co-localization and pathway into consideration. The network shown in Figure-4.12 in-

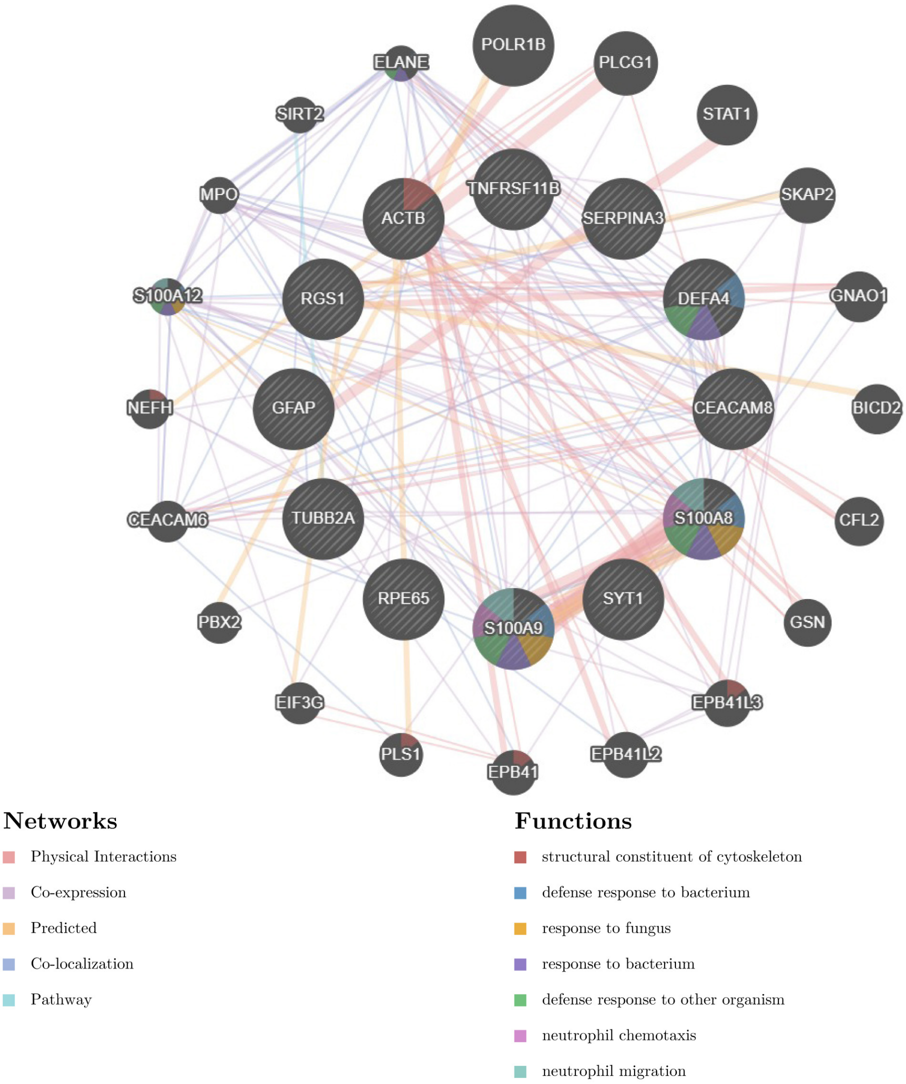

Figure 4.12: Cluster network with overlapping DEGs between AD and other selected pathologies obtained using the online tool GeneMania. Nodes indicate DEGs and links represent functional associations. The node size indicates the rank of the gene considering its association with other nodes and width of the edges represent the percentile contribution of the connecting nodes to a particular functional association.

dicates 32 related genes (nodes) and 183 links between them. The most significant pathways associated with the chosen pathologies and their percentile contributions are a structural constituent of the cytoskeleton (7.35%), defense response to a bacterium (6.58%), response to fungus (27.27%), response to a bacterium (2.99%), defense response to other organisms (2.66%), neutrophil chemotaxis (8.33%), neutrophil migration (8.33%), chemokine production (6.82%), regulation of inflammatory response (2.84%) and inflammatory response (1.77%).

Semantic Similarity and KEGG Enrichment

The semantic similarity measures for DEGs of the selected disease conditions are represented in a matrix as shown in Fig 4.13. AD06_GSE33000 is associated with two

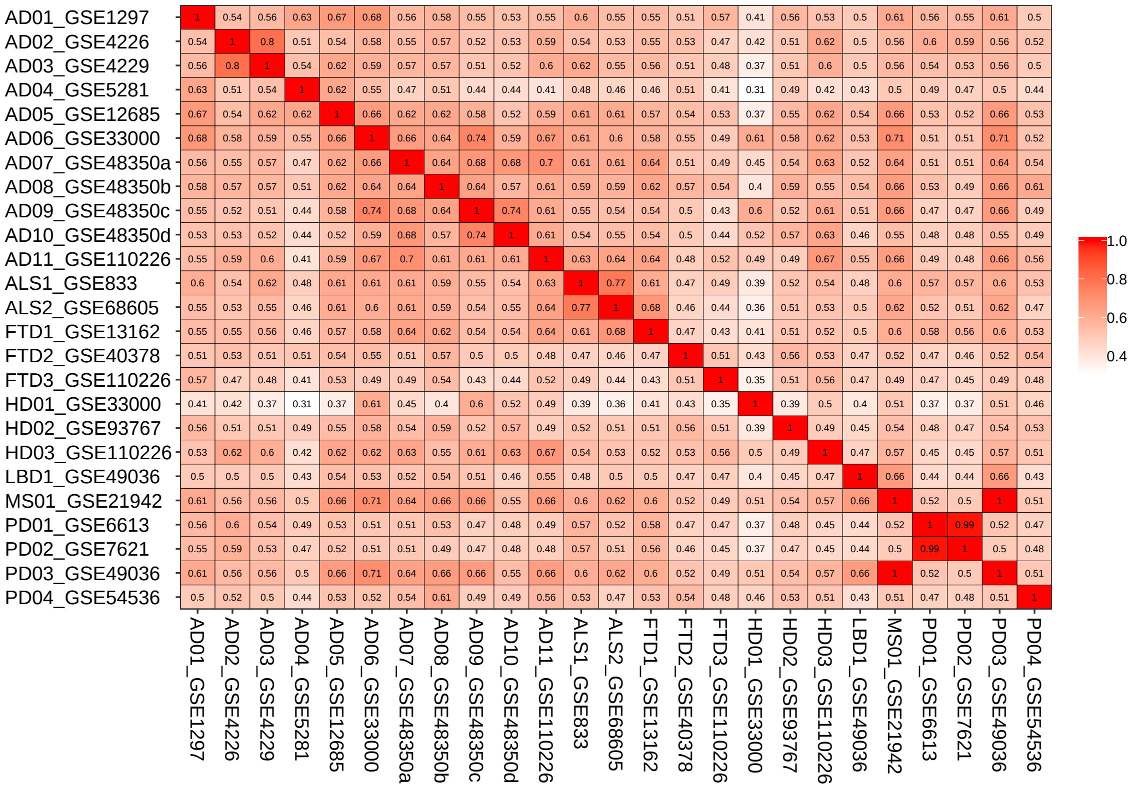

Figure 4.13: Semantic similarity matrix for the differential expressed genes in the five most significant GO terms. The first two letters of each entry represents the selected pathologies (AD-Alzheimer's disease, ALS-Amyotrophic lateral sclerosis, FTD-Frontotemporal dementia, HD-Huntington's disease, LBD-Lewy body disease, MS-Multiple sclerosis, PD-Parkinson's disease).

selected comorbidities: Parkinson's disease and multiple sclerosis exhibiting the value of semantic similarity at least 0.7. Considering other evidence from AD11_GSE110226 and AD07_GSE48350a/b, Parkinson's disease, Huntington's disease, amyotrophic lateral sclerosis, frontotemporal dementia, multiple sclerosis and spinal muscular atrophy are closely associated with AD.

Fig 4.14 depicts the semantic similarity matrix for the top five GO terms. Notably, all AD datasets except AD05_GSE12685 are similar (semantic similarity value of 1) to PD01_GSE6613 dataset considering the top five GO terms. In addition, observing the semantic similarity measure being greater than 0.9, AD05_GSE12685 and AD06_GSE33000 are well clustered with both amyotrophic lateral sclerosis datasets. But if we inspect the semantic similarity measure at least 0.8, all Parkinson's disease, Huntington's disease, Lewy body disease, amyotrophic lateral sclerosis, frontotemporal dementia, multiple sclerosis and spinal muscular atrophy employs significant similarity with some of the AD datasets.

Fig 4.15 represents the matrix of DO terms using semantic similarity. Surprisingly, AD exhibited very trivial association with other NDDs considering the DO terms analysis data. Notable significance was observed between spinal muscular atrophy and amyotrophic lateral sclerosis (0.67). On the other hand, Parkinson's disease showed significant association (0.55) with lewy body disorder.

Fig 4.16 shows the KEGG pathway association with all selected datasets. Resulting

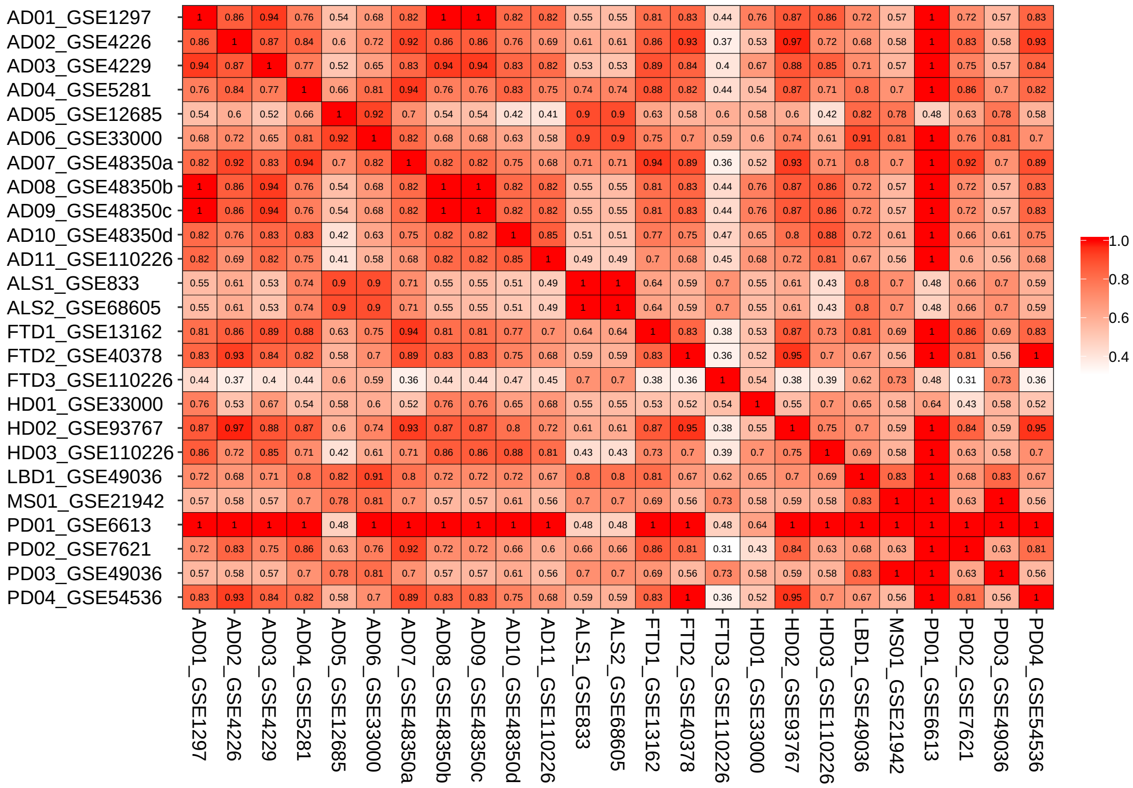

Figure 4.14: Semantic similarity matrix for the five most significant GO terms. Entry names are similar as Fig 4.13.

pathways with at least two occurrences among AD datasets are neuroactive ligand-receptor interaction and malaria. Moreover, recurring pathways common between at least one AD dataset and other pathologies are Parkinson's disease, amphetamine addiction, synaptic vesicle cycle, rheumatoid arthritis, hematopoietic cell lineage, graft-versus-host disease, Staphylococcus aureus infection and IL-17 signaling pathway.

4.4.4 Discussion

In this work, we introduced an analytical framework of bioinformatics analysis for AD-comorbidity studies and demonstrated its efficacy for mining information in public databases. We employed this approach on AD and other NDDs using selected microarray gene expression data from public databases. We applied GSEA to DEGs that we identified, and identified related molecular pathways and their association among selected transcriptomic data using GO and DO. Moreover, we also investigated the effectiveness of semantic similarity as a proximity measure between the diseases using selected ontologies. Identification of the interconnection within a set of pathologies at the molecular level can certainly enrich our insight about the disease mechanism and eventually promotes the possibility for accurate diagnosis and efficacious remedy planning. Our approach leverages publicly available gene expression data from microarray experiments ensuring the possibility of reusing available data. This yields an opportunity to extract hidden information from previously published and publicly accessible datasets. Furthermore, we considered data from different sources and also for different cell types to demonstrate the robustness of the work. Utilization of patient

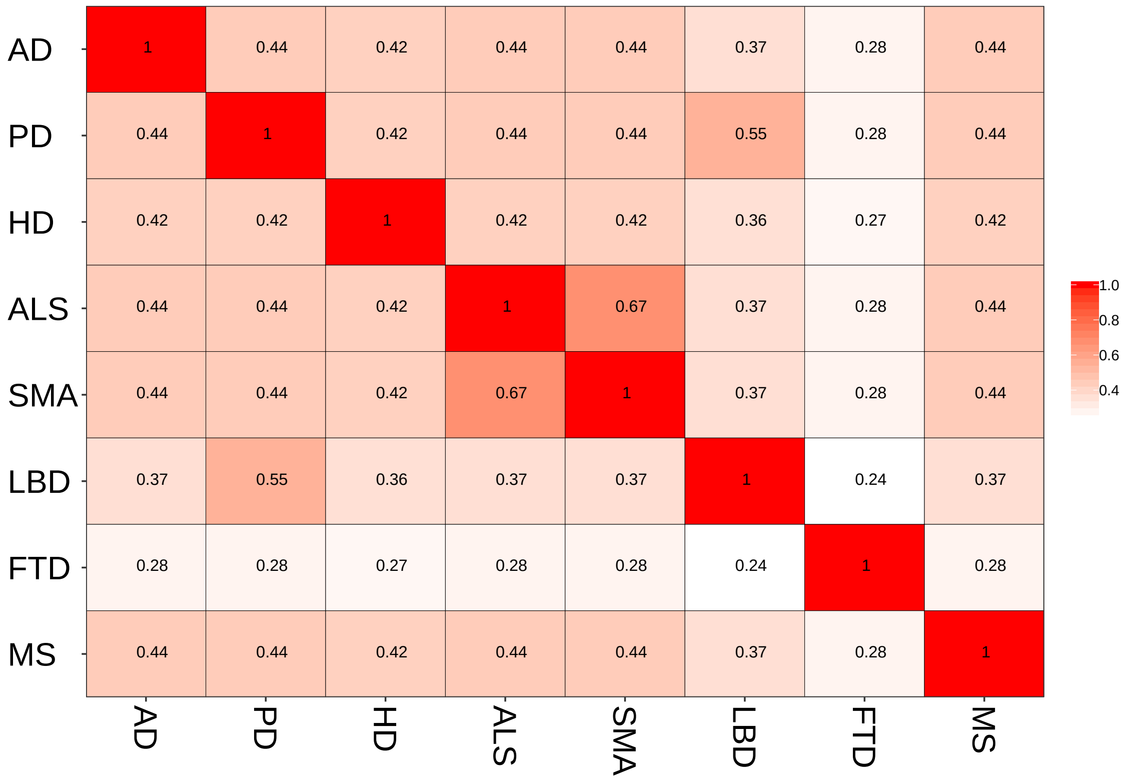

Figure 4.15: Semantic similarity matrix for DO terms. Legends are: AD-Alzheimer's disease, ALS-Amyotrophic lateral sclerosis, FTD-Frontotemporal dementia, HD-Huntington's disease, LBD-Lewy body disease, MS-Multiple sclerosis, PD-Parkinson's disease.

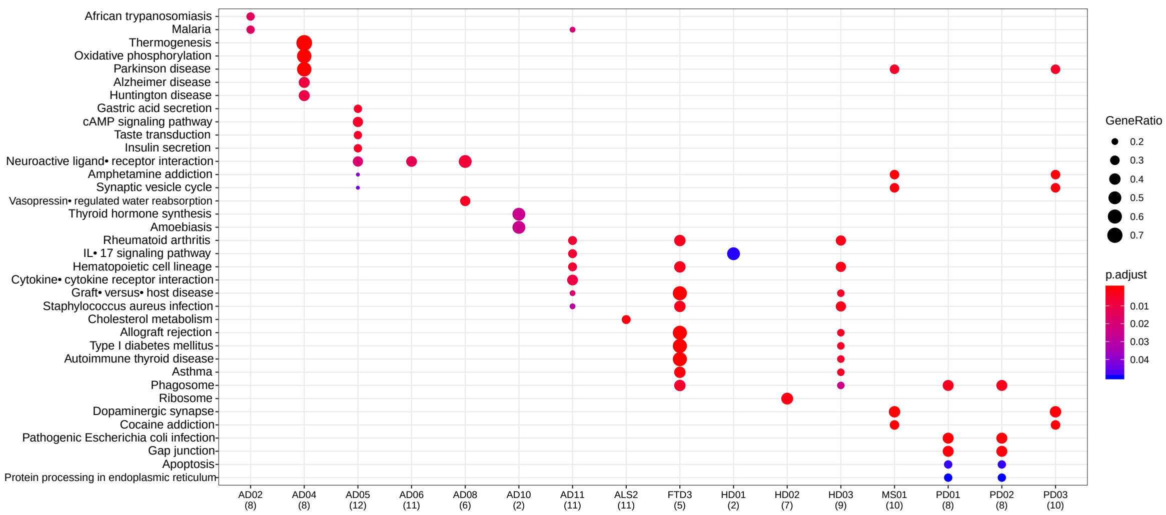

Figure 4.16: KEGG pathway enrichment analysis for differentially expressed genes. Each row represents a KEGG pathway associated with the diseases shown in columns. The domination of genes in the pathway indicated by the dimension of the circles and the range of the circles represents the statistical validation for p-value = 0.05.

omics data is opening new windows for enhancement in clinical decision making including disease risk assessment, accurate diagnosis and subtyping, treatment planning and dose determination [187]. Incorporation of such data into patient care by medical practitioners through clinical activities such as electronic prescribing of medications is a serious prospect. In the near future, aspects of both personalized and preventive medicine will become clinically feasible with potential disease progression assessed by tracking multiple layers of omics and clinical data from healthy individuals. Our work provides methodologies for comorbidity analysis and enhanced visualization as an effective analytical approach that can help professional physicians.

Among the obtained overlapping genes, GFAP has been reported to be associated with AD [188], ALS [189] and MS [190]. Analyzing the co-occurrence of GO terms and molecular pathways between AD and its comorbid neurodegenerative diseases several significant terms and pathways were found to be common. Defects of Oxidative phosphorylation has clear association with AD and PD [191, 192]. Upregulation in cAMP signaling pathway has implication with AD [193]. The association of neuroactive ligand-receptor interaction with α-synuclein is involved in PD [194]. IL-17 signaling pathway has been reported to be involved in the pathogenesis of chronic neuroinflammatory disorder like AD, MS, FTD and HD [195, 196]. The dopaminergic system contributes in neuromodulation and hence the dopaminergic synapse pathways evoke the onset and progression of disorders of central nervous system [197]. The gap junctions connect the cytoplasm of adjacent cells and such interconnections in central nervous system cells maintain normal function. Gap junctions are involved in the pathology of most neurological diseases [198].

We carried out analytical processes for AD and common neurodegenerative comorbidities, although this can be employed for any other AD datasets with other comorbidities if the datasets contain adequate samples for both diseases affected cases and healthy controls. We selected the cutoff sample size 10 considering at least five individuals with active disease state and at least five healthy samples. Our methodology is implemented in an R programming platform that incorporates several other packages from the Bioconductor repository, although these can be easily substituted with another implementation using a different platform. From the methodological point of view, such approaches have been successfully demonstrated various disease interactions recently [69, 199]. It's noteworthy, however, that the dataset selection would have some qualitative and quantitative effects on the outcomes. The findings documented here could be enhanced by incorporating more datasets from other sources as well as different cell types. Nevertheless, our study has employed a new and innovative analytical approach for comorbidity analysis of these complex diseases.

4.5 Experiment Name: Machine learning and bioinformatics models to identify prognostic biomarker and signature genes related to chemoresistance for glioblastoma multiforme

4.5.1 Short summary

Glioblastoma mulitforme (GBM) is the most progressive and lethal form of malignant brain tumor. The genetic mutation is the main reason for this fatal disease. Henceforth, it is essential to identify the causal genetic targets associated with GBM survival. Plenty of publicly accessible gene expression and clinical data for GBM patients from the Gene expression omnibus and the Broad Institute Cancer Genome Atlas (TCGA) dataset can allow us to study patient fatality prediction and thus to identify new GBM biomarkers. In this study we applied bioinformatics and network-based approach to identify the altered genes associated with the GBM comparing with normal mRNA expression data from the brain tissues. Total of 325 genes were found as differentially expressed in GBM. Gene set enrichment analysis through protein-protein interaction (PPI), gene ontology (GO) and KEGG pathways also revealed their significance. We then applied machine-learning approach incorporating Cox Proportional Hazard models to the both clinical and RNA-Seq datasets to determine target biomarkers that affect the survival of the GBM patients. Three genes (PHLDA1, IQGAP and, SPARC) were identified by using univariate approach that have a significant effect on the GBM survival. While treating GBM, chemotherapy is commonly used as the first-line treatment. In some cases, chemotherapy follows the tumor surgery starting with or after the radio therapy to GBM patients. But their progression free survival (PFS) period can be dramatically decreased by the reduced responsiveness to the therapy, usually known as chemoresistance (CR). To identify the candidate signature genes in GBM-chemoresistance (CR), we used the feature importance of Gradient boosting decision tree (GBDT) on the TCGA data. We further evaluated the significance of the identified gens by incorporating support vector machine (SVM). Thus, our combined machine Learning and bioinformatics approach revealed the target signature genes involved in GBM survival and GBM-CR that could be useful to develop potential drug targets for the GBM.

4.5.2 Materials and methods

Data

The microarray gene expression datasets for GBM used in this study were collected from the National Center for Biotechnology Information Gene Expression Omnibus (NCBI-GEO). The datasets are GSE12657, GSE30563 and GSE50161 [200]. GSE12657 is generated by profiling the gene expression of various types of gliomas using Affymetrix microarray. GSE30563 is a comparative study of gene expression data between human brain tumor (GBM) and normal brain tissues through RNA extraction and hybridization on Affymetrix microarrays. GSE50161 is obtained by gene expression profiling of 130 samples for brain tumor and normal brain tissue using Affymetrix HG-U133plus2 chips. The clinical data of GBM patients (Glioblastoma Multiforme TCGA, Provisional-2018) was collected from TCGA website along with the RNA-seq dataset.

The dataset contains 606 cases and 67 clinical features. In clinical datasets we considered some factors such as Sample ID, Race Category, Diagnosis Age, Tumor Disease Anatomic Site, Patient ID, Primary Tumor Site, Censor Status, Overall Survival in Months. The TCGA data contains expression of 20,471 genes and the PFS month for each sample. The dataset includes 111 GBM-CR and 23 GBM-CS (chemosensitive) subjects. We classified the patients whose PFS are at most 9 months as GBM-CR and those whose PFS are at least 15 months as GBM-CS class as proposed by Chen, Kexin, et al. [201].

Analytical Approach

In our present study, we identified most influential genes of GBM survival by analyzing associated genes with clinical information using Cox's Proportional Hazard regression model. For this, at first we identified differential expression genes (DEGs) and modelled their survival function individually to filter out the genes with different expression levels between altered and normal (non-altered) group. Lastly we employed univariate Cox's PH regression model to estimate their significance in GBM survival. We also integrated protein-protein interaction (PPI) and functional enrichment analysis. In order to identify the signature genes associated with GBM-CR, we first normalized the TCGA data using min-max normalization and over-sampled the data using Synthetic Minority Oversampling Technique (SMOTE) algorithm. The preprocessed 222 samples (111 GBM-CR and 111 GBM-CS) were then divided into training (156 samples) and testing (66 samples) samples using holdout method (70%:30%). We used the feature importance of GBDT modeling to calculate average feature importance over 20 iterations to identify the most significant individual signature genes. The frequency of individual gene over 20 iterations were calculated using Equation-3.10. The significance of the selected genes were evaluated using SVM and intense literature review. Figure 4.17 shows the schematic diagram of the methodology.

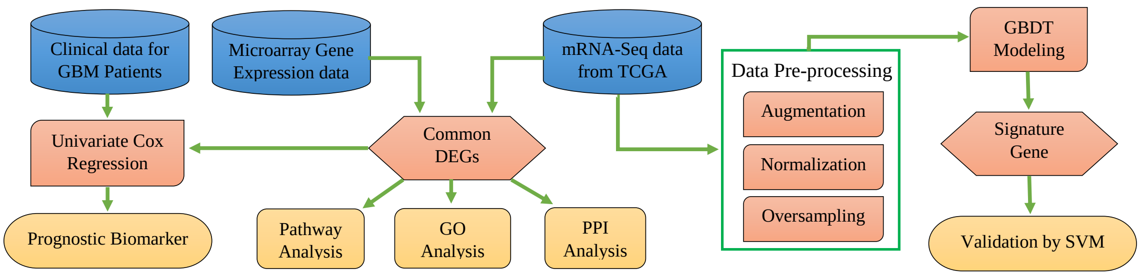

Figure 4.17: Block diagram for the multi-stage methodology of the study.

4.5.3 Result

Differentially Expressed Genes (DEGs) Identification

The three selected microarray datasets were analyzed by comparing normal samples with GBM cases. We identified 325 overlapping DEGs among three datasets (Figure 4.18A for up-regulated and 4.18B for downregulated genes). We also compared the common DEGs with TCGA RNA-seq datasets and found 312 overlapping genes.

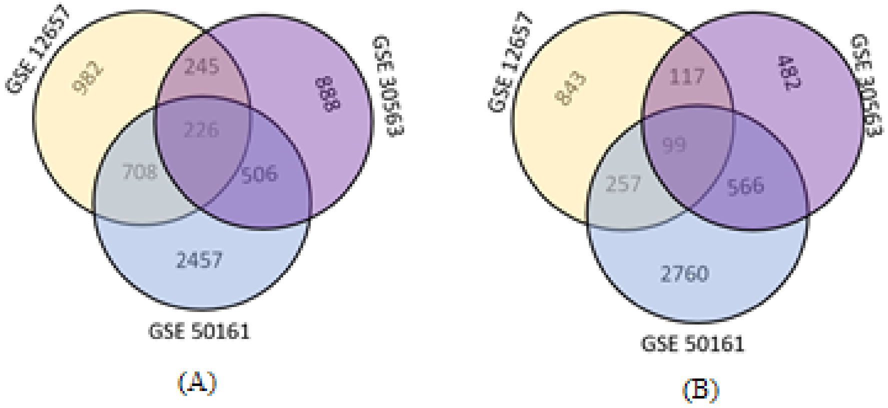

Figure 4.18: Venn diagram showing overlapping a) up-regullated and b) down-regulated DEGs among three microarray datasets.

Functional Enrichment Analysis

The structural enrichment study of the Biological process (BP) GO and KEGG pathway was carried out to show the functions and molecular pathway associated with GBM using 325 common DEGs. Total 20 BP GO terms were found with $p < 0.05$ and the top 10 GO terms are tabulated in Table 4.14. Table 4.15 shows 13 KEGG pathways associated with GBM and central nervous system among total 129 significant pathways ($p < 0.05$).

Analysis of the PPI network

We constructed the PPI network for the 325 overlapping DEGs. Figure 4.19A depicts the network where 264 nodes each representing a protein and 1509 edges each indicating an interaction between each node pair. 11 modules were identified by further analyzing the PPI network using MCODE and the top 2 modules are shown in Figure 4.19B and C. 21 genes belong to module 1 while module 2 contains 20 genes. We also used the CytoHubba for topological analysis of the PPI network and found 10 hub genes (SNAP25, SYN2, RAB3, SYT1, SLC17A7, VAMP2, STXBP1, UNC13A, BSN, and DLG4) as shown in Figure 4.19D.

Identification of survival DEGs

In the survival analysis we considered the overlapping 312 DEGs among the GBM microarray datasets and the TCGA RNA-seq dataset. A predictive model was built by using Univariate Cox proportional hazards stepwise regression. We observed that three genes (PHLDA1, IQGAP2 and SPARC) were significant in the patients' survival having $P < 0.05$ as tabulated in Table 4.16. The survival patters of altered and normal (non-altered) group of these three genes are shown in Figure 4.20. For all the cases, the altered group is less likely to survive compared to the normal group.

GBM-CR related individual signature gene identification

We constructed GBDT model to generate feature importance for each genes considering their gene expression data. We carried out 20 experiments and calculated the frequency

Table 4.14: Top 10 significant Biological process GO terms associated with GBM.

BP GO Term	Adj.P-val	Genes on the pathway
Anterograde trans-synaptic signaling	1.99E-13	SNAP25, DOC2A, CPNE6, GRM1, GRM3, GRIN2A, NPTX1, BSN, GABRD, SCN1B, GABRA2, GABBR2, GABRA1, SLC12A5, SYT1, GAD2, SYN2, GABRG2, CACNB2, CACNB3, CACNB4, DLG4, KCNQ2, AMPH, SCN2B
Chemical synaptic transmission	2.90E-13	SNAP25, DOC2A, CPNE6, GRM1, GRM3, GRIN2A, SLC17A7, NPTX1, BSN, GABRD, SCN1B, GABRA2, GABBR2, GABRA1, UNC13A, SLC12A5, SYT1, GAD2, SYN2, GABRG2, CACNB2, CACNB3, CACNB4, DLG4, KCNQ2, AMPH, SCN2B
Regulation of neuronal synaptic plasticity	1.05E-07	CAMK2B, SYNGR1, RAB3A, DLG4, CAMK2A, PPFIA3, SLC8A2
Positive regulation of synaptic transmission	1.41E-07	MPP2, GRIN2A, RIMS3, SYT1, CLSTN3, DLG4, PTK2B, SHANK2, VAMP2, SLC8A2
Regulation of cardiac conduction	1.41E-07	RYR2, CAV1, ATP1A3, ATP2B3, ATP2B2, CALM3, ATP1B1, CALM1, CALM2, SLC8A2
Regulated exocytosis	1.53E-07	LGALS3BP, SPARC, PROS1, STXBP1, SERPINE1, ANXA5, SYNGR3, TUBA4A, SYNGR1, PPP3CB, RIMS3, TIMP1, CALM1, VAMP2
Positive regulation of calcium ion transmembrane transporter activity	1.87E-07	RYR2, AKAP6, CALM3, ATP1B1, CALM1, CALM2
Cellular response to interferon-gamma	4.62E-07	CAMK2B, SYNCRIP, SP100, HLA-DRB4, CAMK2A, CASP1, CD58, IFI30, GBP2, HLA-G, CAMK2G, CD44
Regulation of dendritic spine development	6.66E-07	CAMK2B, ARHGAP44, NEURL1, CPEB3, SHANK2, LZTS3, CDK5R1
Response to calcium ion	6.73E-07	RYR2, NEUROD2, CLIC4, DMTN, SYT1, CAV1, CALM3, ADCY1, CALM1, CALM2

of each gene. Applying the frequency threshold as 10 we obtained 19 significant signature genes. Figure 4.21 shows the frequency of individual genes over 20 experiments. We validated each gene by modeling SVM classifier and measured evaluation matices accuracy, precision and AUC. Figure 4.22 depicts the indicators when we increase the number of individual genes in each iteration. The 19 significant signature genes showed sufficient classification performance resulting the accuracy, precision and AUC as 0.9682, 0.9583 and 0.9875, respectively. These 19 selected individual signature genes are: PSG1, CLLU1OS, RCOR3, HLA-DRB5, DGCR11, SFTPA2, KRT2, WDR72, COCH, FAM104A, HOXD10, ACHE, OR2B2, TCL1B, CCDC144NL, ULBP2, HELB, CDK15 and PTPN12.

4.5.4 Discussion

Significant efforts have been made in recent decades to enhance the health performance of GBM patients. In this study, we found 325 DEGs (99 down-regulated and 226 up-regulated) to be common in three datasets among which 312 were common with TCGA RNA-seq dataset. Topological analysis of the PPI network using the overlapping DEGs revealed 10 hub genes (SNAP25, SYN2, RAB3, SYT1, SLC17A7, VAMP2,

Table 4.15: Significant KEGG pathways associated with GBM and central nervous system.

KEGG pathway	Adj.P-val	Genes on the pathway
Long-term potentiation	1.70E-13	PRKCG, CAMK2B, PRKCB, CAMK2A, ADCY1, GRM1, GRIN1, PPP3CB, GRIN2A, PPP1R1A, CAMK4, CALM3, CALM1, CAMK2G, CALM2
GABAergic synapse	1.36E-11	NSF, PRKCG, GABRA2, GABBR2, GABRA1, KCNJ6, SLC12A5, PRKCB, GLS2, GAD2, ADCY1, GNG12, GABRG2, GNB5, GABRD
Dopaminergic synapse	4.45E-10	PRKCG, CAMK2B, KCNJ6, PRKCB, KCNJ9, CAMK2A, GNG12, MAPK9, GRIN2A, PPP3CB, KIF5C, CALM3, GNB5, CALM1, CALM2, CAMK2G, KCNJ3
Synaptic vesicle cycle	3.70E-10	NSF, SNAP25, RAB3A, UNC13A, ATP6V1G2, SYT1, STXBP1, CPLX2, DNM1, DNM3, SLC17A7, STX1A, VAMP2
Glutamatergic synapse	5.15E-10	PRKCG, PRKCB, GLS2, ADCY1, GNG12, GRM1, GRIN1, GRM3, GRIN2A, PPP3CB, DLG4, SLC17A7, GNB5, SHANK2, KCNJ3
Retrograde endocannabinoid signaling	1.98E-08	PRKCG, GABRA2, GABRA1, KCNJ6, PRKCB, KCNJ9, ADCY1, GNG12, GABRG2, GRM1, MAPK9, SLC17A7, GNB5, GABRD, KCNJ3
Glioma	3.42E-08	PRKCG, CAMK2B, PRKCB, GADD45A, CAMK4, CAMK2A, PIK3R2, CALM3, CALM1, CAMK2G, CALM2
Cholinergic synapse	3.48E-08	PRKCG, CAMK2B, KCNJ6, PRKCB, CAMK2A, PIK3R2, ADCY1, GNG12, CAMK4, KCNQ2, GNB5, CAMK2G, KCNJ3
Pathways in cancer	6.32E-07	ITGB1, CAMK2B, HDAC1, CAMK2A, CXCR4, PIK3R2, ADCY1, RASGRP1, RELA, MAPK9, FGF9, CAMK2G, RUNX1T1, PRKCG, WNT10B, JAG1, PRKCB, GADD45A, GNG12, COL4A2, COL4A1, CALM3, GNB5, CALM1, CALM2, FGFR2
Neurotrophin signaling pathway	3.90E-06	CAMK2B, MAPK9, CAMK4, ARHGDIB, CAMK2A, PIK3R2, CALM3, CALM1, CALM2, CAMK2G, RELA
Amyotrophic lateral sclerosis (ALS)	1.76E-05	PPP3CB, GRIN2A, CASP1, NEFL, NEFM, NEFH, GRIN1
Serotonergic synapse	2.49E-03	PRKCG, KCNJ6, PRKCB, KCNJ9, GNB5, GNG12, KCNJ3
Alzheimer disease	6.99E-03	GRIN2A, PPP3CB, CALM3, APBB1, CALM1, CALM2, GRIN1, CDK5R1

Table 4.16: Significant KEGG pathways associated with GBM and central nervous system.

Gene Symbol	β	HR	P-value
PHLDA1	2.0094	7.46	2.00E-04
IQGAP2	1.5891	4.8995	1.50E-03
SPARC	2.4654	11.7683	3.01E-02

STXBP1, UNC13A, BSN, and DLG4) that affect several pathways. We also extracted 2 modules from the network and the genes forming these modules are considered to have significant influence in GBM progression. We performed GO Biological process and KEGG pathway analysis to construct molecular function of the common DEGs. Using a log-rank test for each gene, we used univariate Cox PH ratio analysis on mRNA expression results to calculate the survival curve using product limit technique, and determined whether there is any statistically significant difference between the altered

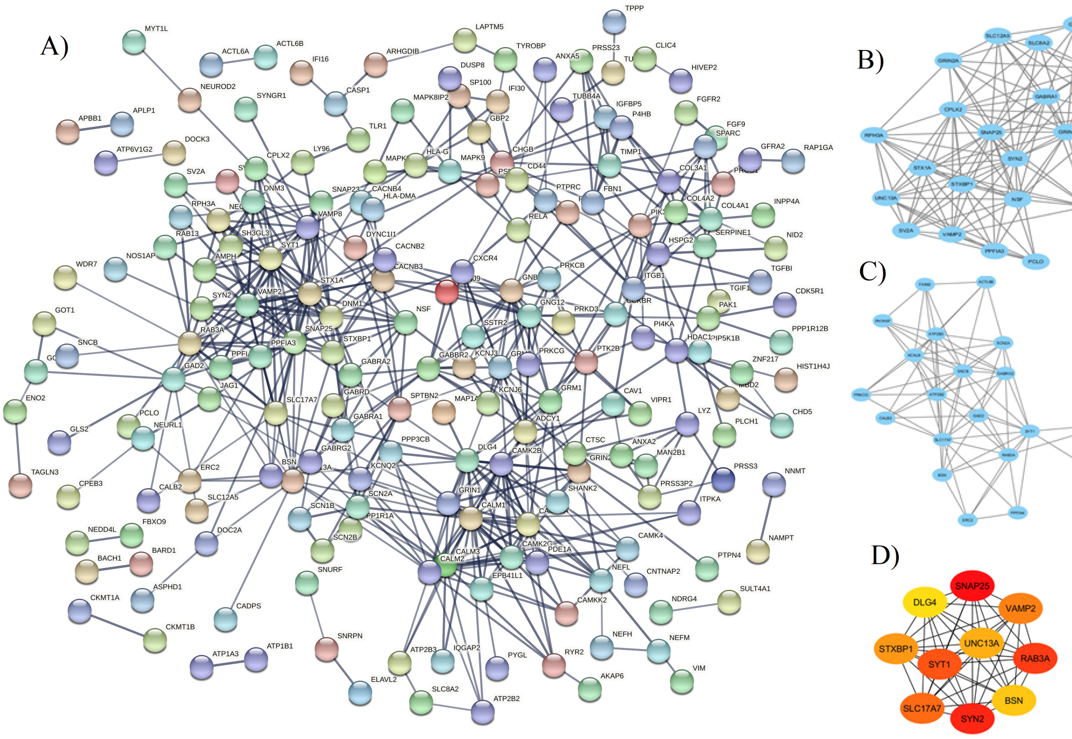

Figure 4.19: A) Protein-protein interaction network for the common DEGs between the three GBM datasets. B) Subnetwork1. C) Subnetwork 2. D) 10 hub genes.

and unaltered classes. In univariate, PHLDA1, IQGAP2, and SPARC are identified four significant genes which are biomarkers of GBM patient. IQGAP2 is a potential biomarker [202]. Due to research, it is not validated in this paper. The miRNA negative correlation with strongest IQGAP was subordinated by IQGAP2. In this present study we saw that the gene was playing vital role in GBM. Shi, Q., et al. showed the reduced tumor cell survival and invasion with SPARC expression with short interfering RNA (siRNA) in glioma cells [203]. AKT and two cytoplasmic kinases, focal adhesion kinase (FAK) and integrin-linked kinase (ILK) are decreased by SPARC siRNA. The SPARC gene is very dangerous and survival probability is very less (Figure 4.20).

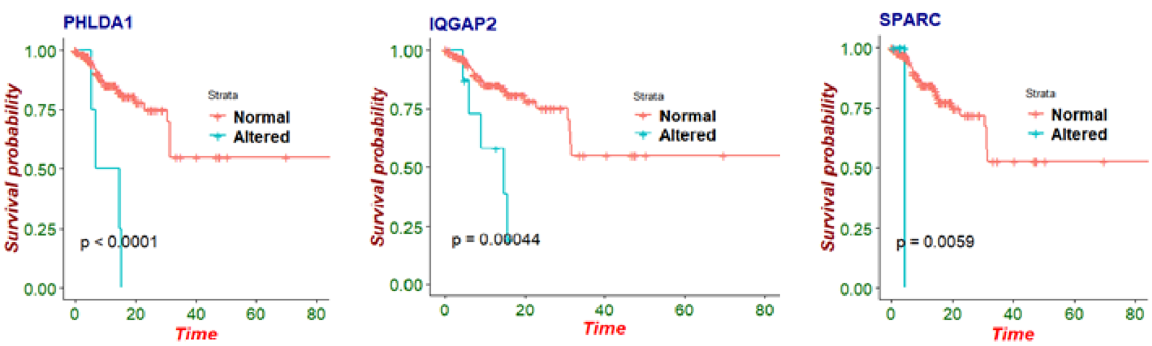

Figure 4.20: Survival pattern comparison of altered and unaltered expression groups for PHLDA1, IQGAP2, and SPARC.

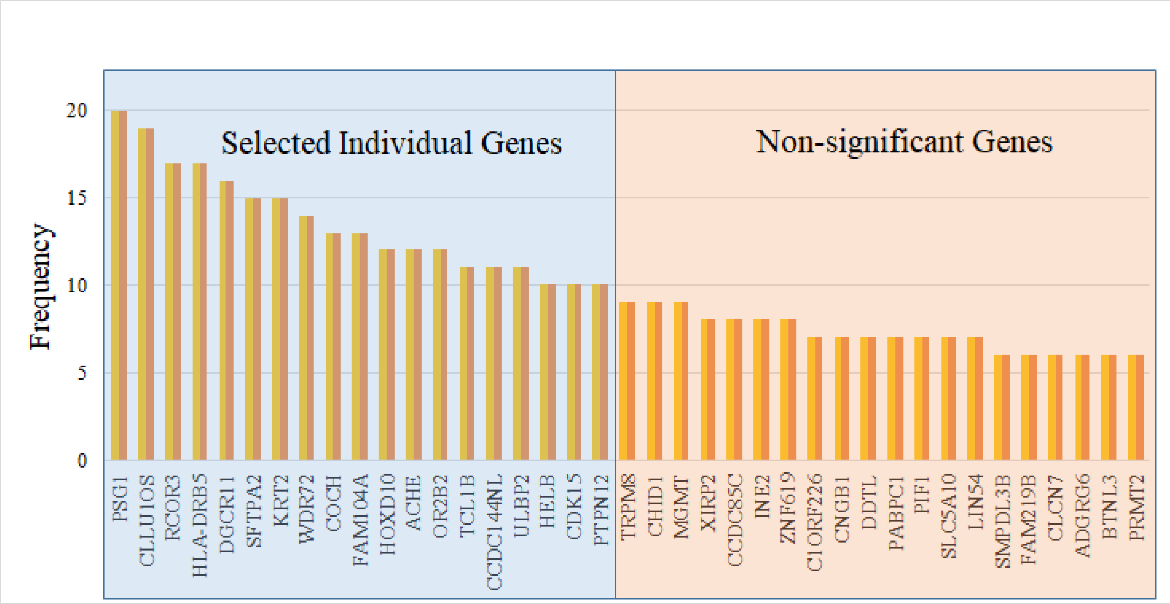

Figure 4.21: The frequency of the selected 19 significant signature gene over 20 experiments.

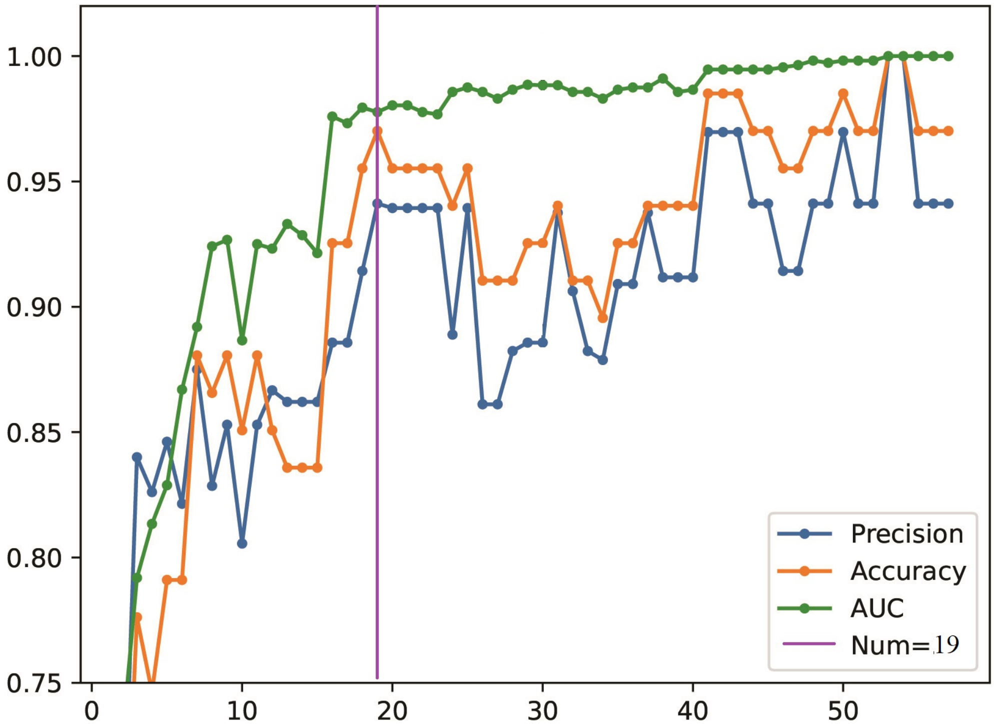

Figure 4.22: The accuracy, precision and AUC indicators while increasing the number of signature genes in each iteration.

Chapter 5

CONCLUSION

5.1 Introduction

Numerous research approaches throughout the last century enriched our understanding about the factors that influence different ND progression and reported a significant number of molecular signatures and therapeutic agents. Still, they could not successfully find any preventive measure, early diagnostic procedure, cure or disease-modifying drug to obstruct or treat these NDs effectively. Besides, inadequate understanding of NDs and their consequences particularly with other neurodegenerative diseases, that means how these diseases influence each other is also unknown. On the other hand, plenty of publicly accessible transcriptomic resources including gene expression data, and bioinformatics and computational tools opened up new opportunity to identify gene pathways that enable disease processes or to influence each other. This dissertation manifests the adaptation of these procurable resources as an attempt of advancement to early detect, prevent, alleviate or prolong survival for NDs. This thesis demonstrates four studies, three targeting AD and one focusing GBM. The concluding remarks of those studies are summarised in Table 5.1.

5.2 Future Work

The gained results in the experiments carried out in this thesis provides better insight about the implication of several causative factors on the development, progression and survival of NDs. Consequently, these can lead to improved treatment and diagnosis of the NDs. Future research directions can be as follows:

i The transcriptomic datasets used in all four experiments are microarray datasets. But microarray datasets are known to be more vulnerable to error. Next generation sequence data or even single cell sequence data could be a better choice. In future, these experiments can be carried out using next generation sequence or single cell sequence data.

ii After clinical validation, the findings of the experiments can lead to improved treatment and diagnosis. So, in future, the gained results can be verified via functional studies.

iii Three experiments were done on AD and one on GBM. In future, similar studies can be carried out for other NDs.

iv The experiments, especially the fourth one, enumerates a lot of candidate signature genes which can be further validated adopting state-of-art methodologies to cut down the list for better reasoning.

v Alongside the analytical frameworks adopted in the studies, improved machine learning methodologies can be incorporated for better accuracy or enhanced performance.

Sl. No.	**Experiment Name**	**Concluding Remarks**
1	Network-based identification of genetic factors in ageing, lifestyle and type 2 diabetes that influence to the progression of Alzheimer's disease	i We identified 484 DEGs from AD brain tissue, of which 27 were also seen in the smoking DEGs gene set. AD data also showed 21 DEGs in common with T2D, and 12 with sedentary lifestyle datasets. ii AD shared less than ten DEGs with the other factors, but 3 (HLA-DRB4, IGH and IGHA2) were commonly up-regulated among the AD, T2D and high alcohol consumption datasets; IGHD and IGHG1 were up-regulated among AD, T2D, alcohol and sedentary lifestyle datasets. iii Protein-protein interaction networks identified 10 hub genes: CREBBP, PRKCB, ITGB1, GAD1, GNB5, PPP3CA, CABP1, SMARCA4, SNAP25 and GRIA1. iv Shared signaling pathway were identified that could enhance our understanding about the mechanisms of AD progression.
2	Systems biology and bioinformatics approach to identify gene signatures, pathways and therapeutic targets of Alzheimer's disease	i 20 common up-regulated DEGs and 9 common down-regulated DEGs were found between blood and brain. ii 18 significant genes were identified which were commonly dysregulated between blood and brain. These are HSD17B1, GAS5, RPS5, VKORC1, GLE1, WDR1, RPL12, MORN1, RAD52, SDR39U1, NPHP4, MT1E, SORD, LINC00638, MCM3AP-AS1, GSDMD, RPS9, and GNL2. iii 10 hub genes SST, GAP43, NRXN1, CHRNA4, VSNL1, HGF, WIF1, PAX3, NFASC and DIMT1 were obtained in PPI analysis. iv Significant molecular pathway and gene ontology indicating AD progression were identified.

3	System biology and bioinformatics pipeline to identify comorbidities risk association: neurodegenerative disorder case study	i We identified genes with abnormal expressions that were common to AD and its comorbidities, as well as shared gene ontology terms and molecular pathways. ii 14 overlapping genes are identified between AD and its NDD comorbidities as ACTB, CEACAM8, COX2, DEFA4, GFAP, MALAT1, RGS1, RPE65, SYT1, S100A8, S100A9, SERPINA3, TNFRSF11B and TUBB2A. iii The semantic similarity measures for DEGs of the selected disease identified close association of AD with PD, HD, ALS, FTD, MS and SMA. iv AD is highly associated with PD, HD, LBD, ALS, FTD, MS and SMA considering semantic similarity for top 5 GO terms. v Our methodological pipeline was implemented in the R platform as an open-source package and available at the following link: `https://github.com/unchowdhury/AD_comorbidity`.
4	Machine learning and bioinformatics models to identify prognostic biomarker and signature genes related to chemoresistance for glioblastoma multiforme	i Total of 325 genes were found as differentially expressed in GBM. ii Gene set enrichment analysis through protein-protein interaction (PPI), gene ontology (GO) and KEGG pathways also revealed their significance. iii Three genes (PHLDA1, IQGAP and, SPARC) were identified by using univariate approach that have a significant effect on the GB survival. iv 19 individual signature genes (PSG1, CLLU1OS, RCOR3, HLA-DRB5, DGCR11, SFTPA2, KRT2, WDR72, COCH, FAM104A, HOXD10, ACHE, OR2B2, TCL1B, CCDC144NL, ULBP2, HELB, CDK15 and PTPN12) were identified having association with GBM-CR.

Table 5.1: Concluding remarks regarding the experiments carried out in this thesis.

Bibliography

[1] V. L. Feigin, A. A. Abajobir, K. H. Abate, F. Abd-Allah, A. M. Abdulle, S. F. Abera, G. Y. Abyu, M. B. Ahmed, A. N. Aichour, I. Aichour *et al.*, "Global, regional, and national burden of neurological disorders during 1990–2015: a systematic analysis for the global burden of disease study 2015," *The Lancet Neurology*, vol. 16, no. 11, pp. 877–897, 2017.

[2] C. J. Murray, T. Vos, R. Lozano, M. Naghavi, A. D. Flaxman, C. Michaud, M. Ezzati, K. Shibuya, J. A. Salomon, S. Abdalla *et al.*, "Disability-adjusted life years (dalys) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010," *The lancet*, vol. 380, no. 9859, pp. 2197–2223, 2012.

[3] W. H. Organization, *Neurological disorders: public health challenges.* World Health Organization, 2006.

[4] A. Burns and S. Iliffe, "Alzheimer's disease. bmj 338, b158," 2009.

[5] "Dementia." [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/dementia

[6] "Alzheimer's disease fact sheet." [Online]. Available: https://www.nia.nih.gov/health/alzheimers-disease-fact-sheet

[7] A. Association *et al.*, "2018 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 14, no. 3, pp. 367–429, 2018.

[8] J. L. Cummings and G. Cole, "Alzheimer disease," *Jama*, vol. 287, no. 18, pp. 2335–2338, 2002.

[9] C. G. Lyketsos, M. C. Carrillo, J. M. Ryan, A. S. Khachaturian, P. Trzepacz, J. Amatniek, J. Cedarbaum, R. Brashear, and D. S. Miller, "Neuropsychiatric symptoms in alzheimer's disease," 2011.

[10] M. Grundman, M. J. Pontecorvo, S. P. Salloway, P. M. Doraiswamy, A. S. Fleisher, C. H. Sadowsky, A. K. Nair, A. Siderowf, M. Lu, A. K. Arora *et al.*, "Potential impact of amyloid imaging on diagnosis and intended management in patients with progressive cognitive decline," *Alzheimer Disease & Associated Disorders*, vol. 27, no. 1, pp. 4–15, 2013.

[11] C. M. van Duijn, P. de Knijff, M. Cruts, A. Wehnert, L. M. Havekes, A. Hofman, and C. Van Broeckhoven, "Apolipoprotein e4 allele in a population–based study of early–onset alzheimer's disease," *Nature genetics*, vol. 7, no. 1, pp. 74–78, 1994.

[12] E. H. Corder, A. M. Saunders, W. J. Strittmatter, D. E. Schmechel, P. C. Gaskell, G. Small, A. Roses, J. Haines, and M. A. Pericak-Vance, “Gene dose of apolipoprotein e type 4 allele and the risk of alzheimer’s disease in late onset families,” *Science*, vol. 261, no. 5123, pp. 921–923, 1993.

[13] K. Okuizumi, O. Onodera, H. Tanaka, S. Tsuji, K. Seki, Y. Namba, K. Ikeda, A. M. Saunders, M. A. Pericak-Vance, and A. D. Roses, “Lack of association of very low density lipoprotein receptor gene polymorphism with caucasian alzheimer’s disease,” *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 40, no. 2, pp. 251–254, 1996.

[14] A. S. Henderson and A. F. Jorm, “Definition and epidemiology of dementia: a review,” *Dementia*, vol. 3, pp. 1–68, 2002.

[15] M. Schreiber, T. D. Bird, and D. W. Tsuang, “Alzheimer’s disease genetics,” *Current Behavioral Neuroscience Reports*, vol. 1, no. 4, pp. 191–196, 2014.

[16] J. S. Goldman, S. E. Hahn, J. W. Catania, S. Larusse-Eckert, M. B. Butson, M. Rumbaugh, M. N. Strecker, J. S. Roberts, W. Burke, R. Mayeux *et al.*, “Genetic counseling and testing for alzheimer disease: joint practice guidelines of the american college of medical genetics and the national society of genetic counselors,” *Genetics in Medicine*, vol. 13, no. 6, p. 597, 2011.

[17] K. Iqbal, F. Liu, and C.-X. Gong, “Alzheimer disease therapeutics: focus on the disease and not just plaques and tangles,” *Biochemical pharmacology*, vol. 88, no. 4, pp. 631–639, 2014.

[18] “Current research areas of alzheimer’s disease.” [Online]. Available: https://alzheimers.com.au/our-work/current-research/

[19] R. M. Young, A. Jamshidi, G. Davis, and J. H. Sherman, “Current trends in the surgical management and treatment of adult glioblastoma,” *Annals of translational medicine*, vol. 3, no. 9, 2015.

[20] C. Wild, B. Stewart, and C. Wild, “World cancer report 2014: World health organization geneva,” 2014.

[21] O. Gallego, “Nonsurgical treatment of recurrent glioblastoma,” *Current oncology*, vol. 22, no. 4, p. e273, 2015.

[22] H. Ohgaki and P. Kleihues, “Genetic pathways to primary and secondary glioblastoma,” *The American journal of pathology*, vol. 170, no. 5, pp. 1445–1453, 2007.

[23] Q. T. Ostrom, H. Gittleman, J. Fulop, M. Liu, R. Blanda, C. Kromer, Y. Wolinsky, C. Kruchko, and J. S. Barnholtz-Sloan, “Cbtrus statistical report: primary brain and central nervous system tumors diagnosed in the united states in 2008-2012,” *Neuro-oncology*, vol. 17, no. suppl_4, pp. iv1–iv62, 2015.

[24] B. M. Alexander and T. F. Cloughesy, “Adult glioblastoma,” *Journal of Clinical Oncology*, vol. 35, no. 21, pp. 2402–2409, 2017.

[25] G. M. McKhann, D. Knopman, H. Chertkow, B. Hyman, C. Jack, C. Kawas, W. Klunk, W. Koroshetz, J. Manley, R. Mayeux *et al.*, "The diagnosis of dementia due to alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," *Alzheimers Dement*, vol. 7, no. 3, pp. 263–9, 2011.

[26] M. S. Albert, S. T. DeKosky, D. Dickson, B. Dubois, H. H. Feldman, N. C. Fox, A. Gamst, D. M. Holtzman, W. J. Jagust, R. C. Petersen *et al.*, "The diagnosis of mild cognitive impairment due to alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," *Focus*, vol. 11, no. 1, pp. 96–106, 2013.

[27] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E. M. Stadlan, "Clinical diagnosis of alzheimer's disease: Report of the nincds-adrda work group* under the auspices of department of health and human services task force on alzheimer's disease," *Neurology*, vol. 34, no. 7, pp. 939–939, 1984.

[28] F. Pasquier, "Early diagnosis of dementia: neuropsychology," *Journal of neurology*, vol. 246, no. 1, pp. 6–15, 1999.

[29] M.-C. Chartier-Harlin, F. Crawford, H. Houlden, A. Warren, D. Hughes, L. Fidani, A. Goate, M. Rossor, P. Roques, J. Hardy *et al.*, "Early-onset alzheimer's disease caused by mutations at codon 717 of the β-amyloid precursor protein gene," *Nature*, vol. 353, no. 6347, pp. 844–846, 1991.

[30] E. Levy-Lahad, W. Wasco, P. Poorkaj, D. M. Romano, J. Oshima, W. H. Pettingell, C.-e. Yu, P. D. Jondro, S. D. Schmidt, K. Wang *et al.*, "Candidate gene for the chromosome 1 familial alzheimer's disease locus," *Science*, vol. 269, no. 5226, pp. 973–977, 1995.

[31] S. Sorbi, P. Forleo, A. Tedde, E. Cellini, M. Ciantelli, S. Bagnoli, and B. Nacmias, "Genetic risk factors in familial alzheimer's disease," *Mechanisms of ageing and development*, vol. 122, no. 16, pp. 1951–1960, 2001.

[32] B. Reisberg, "Alzheimer's disease update," 1985.

[33] J. Lindsay, D. Laurin, R. Verreault, R. Hébert, B. Helliwell, G. B. Hill, and I. McDowell, "Risk factors for alzheimer's disease: a prospective analysis from the canadian study of health and aging," *American journal of epidemiology*, vol. 156, no. 5, pp. 445–453, 2002.

[34] M. Kivipelto, T. Ngandu, L. Fratiglioni, M. Viitanen, I. Kåreholt, B. Winblad, E.-L. Helkala, J. Tuomilehto, H. Soininen, and A. Nissinen, "Obesity and vascular risk factors at midlife and the risk of dementia and alzheimer disease," *Archives of neurology*, vol. 62, no. 10, pp. 1556–1560, 2005.

[35] R. Mayeux and Y. Stern, "Epidemiology of alzheimer disease," *Cold Spring Harbor perspectives in medicine*, p. a006239, 2012.

[36] J. Janson, T. Laedtke, J. E. Parisi, P. O'Brien, R. C. Petersen, and P. C. Butler, "Increased risk of type 2 diabetes in alzheimer disease," *Diabetes*, vol. 53, no. 2, pp. 474–481, 2004.

[37] M. C. Morris, D. A. Evans, J. L. Bienias, C. C. Tangney, D. A. Bennett, N. Aggarwal, J. Schneider, and R. S. Wilson, "Dietary fats and the risk of incident alzheimer disease," *Archives of neurology*, vol. 60, no. 2, pp. 194–200, 2003.

[38] K. J. Anstey, H. A. Mack, and N. Cherbuin, "Alcohol consumption as a risk factor for dementia and cognitive decline: meta-analysis of prospective studies," *The American journal of Geriatric psychiatry*, vol. 17, no. 7, pp. 542–555, 2009.

[39] K. Bettens, K. Sleegers, and C. Van Broeckhoven, "Genetic insights in alzheimer's disease," *The Lancet Neurology*, vol. 12, no. 1, pp. 92–104, 2013.

[40] L. Bertram and R. E. Tanzi, "The genetics of alzheimer's disease," *Progress in molecular biology and translational science*, vol. 107, pp. 79–100, 2012.

[41] R. Guerreiro and J. Hardy, "Genetics of alzheimer's disease," *Neurotherapeutics*, vol. 11, no. 4, pp. 732–737, 2014.

[42] D. W. Dickson, H. Braak, J. E. Duda, C. Duyckaerts, T. Gasser, G. M. Halliday, J. Hardy, J. B. Leverenz, K. Del Tredici, Z. K. Wszolek *et al.*, "Neuropathological assessment of parkinson's disease: refining the diagnostic criteria," *The Lancet Neurology*, vol. 8, no. 12, pp. 1150–1157, 2009.

[43] G. M. Halliday, J. L. Holton, T. Revesz, and D. W. Dickson, "Neuropathology underlying clinical variability in patients with synucleinopathies," *Acta neuropathologica*, vol. 122, no. 2, pp. 187–204, 2011.

[44] W. Poewe, K. Seppi, C. M. Tanner, G. M. Halliday, P. Brundin, J. Volkmann, A.-E. Schrag, and A. E. Lang, "Parkinson disease," *Nature reviews Disease primers*, vol. 3, no. 1, pp. 1–21, 2017.

[45] M. H. Yan, X. Wang, and X. Zhu, "Mitochondrial defects and oxidative stress in alzheimer disease and parkinson disease," *Free Radical Biology and Medicine*, vol. 62, pp. 90–101, 2013.

[46] J. M. Shoffner, M. D. Brown, A. Torroni, M. T. Lott, M. F. Cabell, S. S. Mirra, M. F. Beal, C.-C. Yang, M. Gearing, R. Salvo *et al.*, "Mitochondrial dna variants observed in alzheimer disease and parkinson disease patients," *Genomics*, vol. 17, no. 1, pp. 171–184, 1993.

[47] F. O. Walker, "Huntington's disease," *The Lancet*, vol. 369, no. 9557, pp. 218–228, 2007.

[48] L. S. Forno, "Neuropathologic features of parkinson's, huntington's, and alzheimer's diseases." *Annals of the New York Academy of Sciences*, vol. 648, pp. 6–16, 1992.

[49] M. C. Kiernan, S. Vucic, B. C. Cheah, M. R. Turner, A. Eisen, O. Hardiman, J. R. Burrell, and M. C. Zoing, "Amyotrophic lateral sclerosis," *The lancet*, vol. 377, no. 9769, pp. 942–955, 2011.

[50] D. Majoor-Krakauer, R. Ottman, W. G. Johnson, and L. P. Rowland, "Familial aggregation of amyotrophic lateral sclerosis, dementia, and parkinson's disease: evidence of shared genetic susceptibility," *Neurology*, vol. 44, no. 10, pp. 1872–1872, 1994.

[51] M. Mhatre, R. A. Floyd, and K. Hensley, "Oxidative stress and neuroinflammation in alzheimer's disease and amyotrophic lateral sclerosis: common links and potential therapeutic targets," *Journal of Alzheimer's disease*, vol. 6, no. 2, pp. 147–157, 2004.

[52] "Huntington's disease information page." [Online]. Available: https://www.ninds.nih.gov/Disorders/All-Disorders/Huntingtons-Disease-Information-Page

[53] M. R. Lunn and C. H. Wang, "Spinal muscular atrophy," *The Lancet*, vol. 371, no. 9630, pp. 2120–2133, 2008.

[54] M. G. Spillantini, M. L. Schmidt, V. M.-Y. Lee, J. Q. Trojanowski, R. Jakes, and M. Goedert, "α-synuclein in lewy bodies," *Nature*, vol. 388, no. 6645, pp. 839–840, 1997.

[55] R. D. Terry, E. Masliah, D. P. Salmon, N. Butters, R. DeTeresa, R. Hill, L. A. Hansen, and R. Katzman, "Physical basis of cognitive alterations in alzheimer's disease: synapse loss is the major correlate of cognitive impairment," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 30, no. 4, pp. 572–580, 1991.

[56] S. T. DeKosky and S. W. Scheff, "Synapse loss in frontal cortex biopsies in alzheimer's disease: correlation with cognitive severity," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 27, no. 5, pp. 457–464, 1990.

[57] M. Hashimoto and E. Masliah, "Alpha-synuclein in lewy body disease and alzheimer's disease," *Brain pathology*, vol. 9, no. 4, pp. 707–720, 1999.

[58] J. S. Snowden, D. Neary, and D. M. Mann, "Frontotemporal dementia," *The British journal of psychiatry*, vol. 180, no. 2, pp. 140–143, 2002.

[59] C. L. Stopford, J. C. Thompson, D. Neary, A. M. Richardson, and J. S. Snowden, "Working memory, attention, and executive function in alzheimer's disease and frontotemporal dementia," *Cortex*, vol. 48, no. 4, pp. 429–446, 2012.

[60] M. Sospedra and R. Martin, "Immunology of multiple sclerosis," *Annu. Rev. Immunol.*, vol. 23, pp. 683–747, 2005.

[61] A. Dal Bianco, M. Bradl, J. Frischer, A. Kutzelnigg, K. Jellinger, and H. Lassmann, "Multiple sclerosis and alzheimer's disease," *Annals of neurology*, vol. 63, no. 2, pp. 174–183, 2008.

[62] M. A. Moni and P. Liò, "comor: a software for disease comorbidity risk assessment," *Journal of clinical bioinformatics*, vol. 4, no. 1, p. 8, 2014.

[63] ——, "How to build personalized multi-omics comorbidity profiles," *Frontiers in cell and developmental biology*, vol. 3, p. 28, 2015.

[64] M. A. Moni, H. Xu, and P. Lio, "Cytocom: a cytoscape app to visualize, query and analyse disease comorbidity networks," *Bioinformatics*, vol. 31, no. 6, pp. 969–971, 2014.

[65] A. Gutiérrez-Sacristán, À. Bravo, A. Giannoula, M. A. Mayer, F. Sanz, and L. I. Furlong, "comorbidity: an r package for the systematic analysis of disease comorbidities," *Bioinformatics*, vol. 34, no. 18, pp. 3228–3230, 2018.

[66] F. Ronzano, A. Gutiérrez-Sacristán, and L. I. Furlong, "Comorbidity4j: a tool for interactive analysis of disease comorbidities over large patient datasets," *Bioinformatics*, 2019.

[67] E. Capobianco *et al.*, "Comorbidity: a multidimensional approach," *Trends in molecular medicine*, vol. 19, no. 9, pp. 515–521, 2013.

[68] E. Del Prete, A. Facchiano, and P. Liò, "Bioinformatics methodologies for coeliac disease and its comorbidities," *Briefings in bioinformatics*, 2018.

[69] M. H. Rahman, S. Peng, X. Hu, C. Chen, S. Uddin, J. M. Quinn, and M. A. Moni, "Bioinformatics methodologies to identify interactions between type 2 diabetes and neurological comorbidities," *IEEE Access*, vol. 7, pp. 183 948–183 970, 2019.

[70] L. S. Hu, S. Ning, J. M. Eschbacher, N. Gaw, A. C. Dueck, K. A. Smith, P. Nakaji, J. Plasencia, S. Ranjbar, S. J. Price *et al.*, "Multi-parametric mri and texture analysis to visualize spatial histologic heterogeneity and tumor extent in glioblastoma," *PloS one*, vol. 10, no. 11, p. e0141506, 2015.

[71] A. Chaddad, P. Daniel, C. Desrosiers, M. Toews, and B. Abdulkarim, "Novel radiomic features based on joint intensity matrices for predicting glioblastoma patient survival time," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 795–804, 2018.

[72] S. Shukla, J. R. Evans, R. Malik, F. Y. Feng, S. M. Dhanasekaran, X. Cao, G. Chen, D. G. Beer, H. Jiang, and A. M. Chinnaiyan, "Development of a rna-seq based prognostic signature in lung adenocarcinoma," *JNCI: Journal of the National Cancer Institute*, vol. 109, no. 1, 2017.

[73] M. Bredel, "Anticancer drug resistance in primary human brain tumors," *Brain Research Reviews*, vol. 35, no. 2, pp. 161–204, 2001.

[74] P. Lønning and S. Knappskog, "Mapping genetic alterations causing chemoresistance in cancer: identifying the roads by tracking the drivers," *Oncogene*, vol. 32, no. 46, pp. 5315–5330, 2013.

[75] R. Upstill-Goddard, D. Eccles, J. Fliege, and A. Collins, "Machine learning approaches for the discovery of gene–gene interactions in disease data," *Briefings in bioinformatics*, vol. 14, no. 2, pp. 251–260, 2013.

[76] S. N. Dorman, K. Baranova, J. H. Knoll, B. L. Urquhart, G. Mariani, M. L. Carcangiu, and P. K. Rogan, "Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning," *Molecular oncology*, vol. 10, no. 1, pp. 85–100, 2016.

[77] S. Wani, J. Drahos, M. B. Cook, A. Rastogi, A. Bansal, R. Yen, P. Sharma, and A. Das, "Comparison of endoscopic therapies and surgical resection in patients with early esophageal cancer: a population-based study," *Gastrointestinal endoscopy*, vol. 79, no. 2, pp. 224–232, 2014.

[78] T. Ideker and N. J. Krogan, "Differential network biology," *Molecular systems biology*, vol. 8, no. 1, p. 565, 2012.

[79] A. Sturn, J. Quackenbush, and Z. Trajanoski, "Genesis: cluster analysis of microarray data," *Bioinformatics*, vol. 18, no. 1, pp. 207–208, 2002.

[80] J. N. Hirschhorn and M. J. Daly, "Genome-wide association studies for common diseases and complex traits," *Nature reviews genetics*, vol. 6, no. 2, p. 95, 2005.

[81] A. Rzhetsky, D. Wajngurt, N. Park, and T. Zheng, "Probing genetic overlap among complex human phenotypes," *Proceedings of the National Academy of Sciences*, vol. 104, no. 28, pp. 11 694–11 699, 2007.

[82] M. A. Moni and P. Lio', "Genetic profiling and comorbidities of zika infection," *The Journal of infectious diseases*, vol. 216, no. 6, pp. 703–712, 2017.

[83] J. De Las Rivas and C. Fontanillo, "Protein–protein interactions essentials: key concepts to building and analyzing interactome networks," *PLoS Comput Biol*, vol. 6, no. 6, p. e1000807, 2010.

[84] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork *et al.*, "The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible," *Nucleic acids research*, p. gkw937, 2016.

[85] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[86] S.-H. Chen, C.-H. Chin, H.-H. Wu, C.-W. Ho, M.-T. Ko, and C.-Y. Lin, "cytohubba: A cytoscape plug-in for hub object analysis in network biology," in *20th International Conference on Genome Informatics*, 2009.

[87] G. D. Bader and C. W. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC bioinformatics*, vol. 4, no. 1, p. 2, 2003.

[88] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15 545–15 550, 2005.

[89] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[90] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "Kegg: new perspectives on genomes, pathways, diseases and drugs," *Nucleic acids research*, vol. 45, no. D1, pp. D353–D361, 2017.

[91] E. Wingender, P. Dietze, H. Karas, and R. Knüppel, “Transfac: a database on transcription factors and their dna binding sites,” *Nucleic acids research*, vol. 24, no. 1, pp. 238–241, 1996.

[92] O. Fornes, J. A. Castro-Mondragon, A. Khan, R. Van der Lee, X. Zhang, P. A. Richmond, B. P. Modi, S. Correard, M. Gheorghe, D. Baranašić *et al.*, “Jaspar 2020: update of the open-access database of transcription factor binding profiles,” *Nucleic acids research*, vol. 48, no. D1, pp. D87–D92, 2020.

[93] S.-D. Hsu, F.-M. Lin, W.-Y. Wu, C. Liang, W.-C. Huang, W.-L. Chan, W.-T. Tsai, G.-Z. Chen, C.-J. Lee, C.-M. Chiu *et al.*, “mirtarbase: a database curates experimentally validated microrna–target interactions,” *Nucleic acids research*, vol. 39, no. suppl_1, pp. D163–D169, 2011.

[94] M. Yoo, J. Shin, J. Kim, K. A. Ryall, K. Lee, S. Lee, M. Jeon, J. Kang, and A. C. Tan, “Dsigdb: drug signatures database for gene set analysis,” *Bioinformatics*, vol. 31, no. 18, pp. 3069–3071, 2015.

[95] G. Consortium *et al.*, “Genetic effects on gene expression across human tissues,” *Nature*, vol. 550, no. 7675, pp. 204–213, 2017.

[96] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, “A new method to measure the semantic similarity of go terms,” *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, 2007.

[97] D. R. Cox, “Regression models and life-tables,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[98] S. A. Waldman, T. Hyslop, S. Schulz, A. Barkun, K. Nielsen, J. Haaf, C. Bonaccorso, Y. Li, and D. S. Weinberg, “Association of gucy2c expression in lymph nodes with time to recurrence and disease-free survival in pn0 colorectal cancer,” *Jama*, vol. 301, no. 7, pp. 745–752, 2009.

[99] M. A. Hossain, S. M. S. Islam, J. M. Quinn, F. Huq, and M. A. Moni, “Machine learning and bioinformatics models to identify gene expression patterns of ovarian cancer associated with disease progression and mortality,” *Journal of biomedical informatics*, vol. 100, p. 103313, 2019.

[100] H. K. Rana, M. R. Akhtar, M. B. Islam, M. B. Ahmed, P. Lió, F. Huq, J. M. Quinn, and M. A. Moni, “Machine learning and bioinformatics models to identify pathways that mediate influences of welding fumes on cancer progression,” *Scientific reports*, vol. 10, no. 1, pp. 1–15, 2020.

[101] J. Shumake, T. T. Mallard, J. E. McGeary, and C. G. Beevers, “Inclusion of genetic variants in an ensemble of gradient boosting decision trees does not improve the prediction of citalopram treatment response,” *Scientific reports*, vol. 11, no. 1, pp. 1–9, 2021.

[102] P. Xuan, C. Sun, T. Zhang, Y. Ye, T. Shen, and Y. Dong, “Gradient boosting decision tree-based method for predicting interactions between target genes and drugs,” *Frontiers in genetics*, vol. 10, p. 459, 2019.

[103] C. D. A. Vanitha, D. Devaraj, and M. Venkatesulu, "Gene expression data classification using support vector machine and mutual information-based gene selection," *procedia computer science*, vol. 47, pp. 13–21, 2015.

[104] F. Chu and L. Wang, "Gene expression data analysis using support vector machines," in *Proceedings of the International Joint Conference on Neural Networks, 2003.*, vol. 3. IEEE, 2003, pp. 2268–2271.

[105] E. M. Blalock, J. W. Geddes, K. C. Chen, N. M. Porter, W. R. Markesbery, and P. W. Landfield, "Incipient alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses," *Proceedings of the National Academy of Sciences*, vol. 101, no. 7, pp. 2173–2178, 2004.

[106] R. E. MacLaren, W. Cui, H. Lu, S. Simard, and K. Cianflone, "Association of adipocyte genes with asp expression: a microarray analysis of subcutaneous and omental adipose tissue in morbidly obese subjects," *BMC medical genomics*, vol. 3, no. 1, p. 3, 2010.

[107] U. Raue, T. A. Trappe, S. T. Estrem, H.-R. Qian, L. M. Helvering, R. C. Smith, and S. Trappe, "Transcriptome signature of resistance exercise adaptations: mixed muscle and fiber type specific profiles in young and old adults," *Journal of applied physiology*, vol. 112, no. 10, pp. 1625–1636, 2012.

[108] S. Radom-Aizik, S. Hayek, I. Shahar, G. Rechavi, N. Kaminski, and I. Ben-Dov, "Effects of aerobic training on gene expression in skeletal muscle of elderly men." *Medicine and science in sports and exercise*, vol. 37, no. 10, pp. 1680–1696, 2005.

[109] S. Kakehi, Y. Tamura, K. Takeno, Y. Sakurai, M. Kawaguchi, T. Watanabe, T. Funayama, F. Sato, S.-i. Ikeda, A. Kanazawa *et al.*, "Increased intramyocellular lipid/impaired insulin sensitivity is associated with altered lipid metabolic genes in muscle of high responders to a high-fat diet," *American Journal of Physiology-Endocrinology and Metabolism*, vol. 310, no. 1, pp. E32–E40, 2015.

[110] H. Misu, T. Takamura, H. Takayama, H. Hayashi, N. Matsuzawa-Nagata, S. Kurita, K. Ishikura, H. Ando, Y. Takeshita, T. Ota *et al.*, "A liver-derived secretory protein, selenoprotein p, causes insulin resistance," *Cell metabolism*, vol. 12, no. 5, pp. 483–495, 2010.

[111] F. Herse, R. Dechend, N. K. Harsem, G. Wallukat, J. Janke, F. Qadri, L. Hering, D. N. Muller, F. C. Luft, and A. C. Staff, "Dysregulation of the circulating and tissue-based renin-angiotensin system in preeclampsia," *Hypertension*, vol. 49, no. 3, pp. 604–611, 2007.

[112] D. G. Hebels, K. M. Sveje, M. C. de Kok, M. H. van Herwijnen, G. G. Kuhnle, L. G. Engels, C. B. Vleugels-Simon, W. G. Mares, M. Pierik, A. A. Masclee *et al.*, "N-nitroso compound exposure-associated transcriptomic profiles are indicative of an increased risk for colorectal cancer," *Cancer letters*, vol. 309, no. 1, pp. 1–10, 2011.

[113] J. N. McClintick, A. I. Brooks, L. Deng, L. Liang, J. C. Wang, M. Kapoor, X. Xuei, T. Foroud, J. A. Tischfield, and H. J. Edenberg, "Ethanol treatment of

lymphoblastoid cell lines from alcoholics and non-alcoholics causes many subtle changes in gene expression," *Alcohol*, vol. 48, no. 6, pp. 603–610, 2014.

[114] P. Büttner, S. Mosig, and H. Funke, "Gene expression profiles of t lymphocytes are sensitive to the influence of heavy smoking: a pilot study," *Immunogenetics*, vol. 59, no. 1, pp. 37–43, 2007.

[115] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann *et al.*, "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update," *Nucleic acids research*, vol. 44, no. W1, pp. W90–W97, 2016.

[116] E. Lin, S.-J. Tsai, P.-H. Kuo, Y.-L. Liu, A. C. Yang, and C.-F. Kao, "Association and interaction effects of alzheimer's disease-associated genes and lifestyle on cognitive aging in older adults in a taiwanese population," *Oncotarget*, vol. 8, no. 15, p. 24077, 2017.

[117] L. Shen, Y. Chen, A. Yang, C. Chen, L. Liao, S. Li, M. Ying, J. Tian, Q. Liu, and J. Ni, "Redox proteomic profiling of specifically carbonylated proteins in the serum of triple transgenic alzheimer's disease mice," *International journal of molecular sciences*, vol. 17, no. 4, p. 469, 2016.

[118] C. Rouaux, J.-P. Loeffler, and A.-L. Boutillier, "Targeting creb-binding protein (cbp) loss of function as a therapeutic strategy in neurological disorders," *Biochemical pharmacology*, vol. 68, no. 6, pp. 1157–1164, 2004.

[119] F. Petrif, R. H. Giles, H. G. Dauwerse, J. J. Saris, R. C. Hennekam, M. Masuno, N. Tommerup, G.-J. B. van Ommen, R. H. Goodman, D. J. Peters *et al.*, "Rubinstein-taybi syndrome caused by mutations in the transcriptional co-activator cbp," *Nature*, vol. 376, no. 6538, p. 348, 1995.

[120] W. Tang, F. Yang, W. Lu, J. Ni, J. Zhang, W. Tang, W. Tang, and C. Zhang, "Association study of creb1 and cbp genes with alzheimer's disease in han chinese," *Asia-Pacific Psychiatry*, vol. 9, no. 3, p. e12274, 2017.

[121] M. R. Rahman, T. Islam, B. Turanli, T. Zaman, H. M. Faruquee, M. M. Rahman, M. N. H. Mollah, R. K. Nanda, K. Y. Arga, E. Gov *et al.*, "Network-based approach to identify molecular signatures and therapeutic agents in alzheimer's disease," *Computational biology and chemistry*, vol. 78, pp. 431–439, 2019.

[122] E. Rosa and M. Fahnestock, "Creb expression mediates amyloid β-induced basal bdnf downregulation," *Neurobiology of aging*, vol. 36, no. 8, pp. 2406–2413, 2015.

[123] A. Antonell, A. Lladó, R. Sánchez-Valle, C. Sanfeliu, T. Casserras, L. Rami, C. Muñoz-García, A. Dangla-Valls, M. Balasa, P. Boya *et al.*, "Altered blood gene expression of tumor-related genes (prkcb, becn1, and cdkn2a) in alzheimer's disease," *Molecular neurobiology*, vol. 53, no. 9, pp. 5902–5911, 2016.

[124] S. I. Alfonso, J. A. Callender, B. Hooli, C. E. Antal, K. Mullin, M. A. Sherman, S. E. Lesné, M. Leitges, A. C. Newton, R. E. Tanzi *et al.*, "Gain-of-function mutations in protein kinase cα (pkcα) may promote synaptic defects in alzheimer's disease," *Sci. Signal.*, vol. 9, no. 427, pp. ra47–ra47, 2016.

[125] L. J. Sosa, J. Bergman, A. Estrada-Bernal, T. J. Glorioso, J. M. Kittelson, and K. H. Pfenninger, "Amyloid precursor protein is an autonomous growth cone adhesion molecule engaged in contact guidance," *PLoS One*, vol. 8, no. 5, p. e64521, 2013.

[126] G. Ma, M. Liu, K. Du, X. Zhong, S. Gong, L. Jiao, and M. Wei, "Differential expression of mrnas in the brain tissues of patients with alzheimer's disease based on geo expression profile and its clinical significance," *BioMed research international*, vol. 2019, 2019.

[127] K. D. Siegmund, C. M. Connor, M. Campan, T. I. Long, D. J. Weisenberger, D. Biniszkiewicz, R. Jaenisch, P. W. Laird, and S. Akbarian, "Dna methylation in the human cerebral cortex is dynamically regulated throughout the life span and involves differentiated neurons," *PloS one*, vol. 2, no. 9, p. e895, 2007.

[128] H. E. Shamseldin, I. Masuho, A. Alenizi, S. Alyamani, D. N. Patil, N. Ibrahim, K. A. Martemyanov, and F. S. Alkuraya, "Gnb5 mutation causes a novel neuropsychiatric disorder featuring attention deficit hyperactivity disorder, severely impaired language development and normal cognition," *Genome biology*, vol. 17, no. 1, p. 195, 2016.

[129] M. J. Chiocco, X. Zhu, D. Walther, O. Pletnikova, J. C. Troncoso, G. R. Uhl, and Q.-R. Liu, "Fine mapping of calcineurin (ppp3ca) gene reveals novel alternative splicing patterns, association of 5 utr trinucleotide repeat with addiction vulnerability, and differential isoform expression in alzheimer's disease," *Substance use & misuse*, vol. 45, no. 11, pp. 1809–1826, 2010.

[130] N. Malakooti, C. Fowler, I. Volitakis, C. McLean, R. Kim, A. Bush, A. Rembach, M. Pritchard, D. Finkelstein, P. Adlard *et al.*, "The down syndrome-associated protein, regulator of calcineurin-1, is altered in alzheimer's disease and dementia with lewy bodies," *Journal of Alzheimer's disease & Parkinsonism*, vol. 9, no. 2, 2019.

[131] F. A. Romero, A. M. Taylor, T. D. Crawford, V. Tsui, A. Cote, and S. Magnuson, "Disrupting acetyl-lysine recognition: progress in the development of bromodomain inhibitors," *Journal of medicinal chemistry*, vol. 59, no. 4, pp. 1271–1298, 2015.

[132] M. Porkholm, A. Raunio, R. Vainionpää, T. Salonen, J. Hernesniemi, L. Valanne, J. Satopää, A. Karppinen, M. Oinas, O. Tynninen *et al.*, "Molecular alterations in pediatric brainstem gliomas," *Pediatric blood & cancer*, vol. 65, no. 1, p. e26751, 2018.

[133] S. Karmakar, L. G. Sharma, A. Roy, A. Patel, and L. M. Pandey, "Neuronal snare complex: A protein folding system with intricate protein-protein interactions, and its common neuropathological hallmark, snap25," *Neurochemistry international*, 2018.

[134] E. Bereczki, P. T. Francis, D. Howlett, J. B. Pereira, K. Höglund, A. Bogstedt, A. Cedazo-Minguez, J.-H. Baek, T. Hortobágyi, J. Attems *et al.*, "Synaptic proteins predict cognitive decline in alzheimer's disease and lewy body dementia," *Alzheimer's & Dementia*, vol. 12, no. 11, pp. 1149–1158, 2016.

[135] M. Kowalska, M. Prendecki, W. Kozubski, M. Lianeri, and J. Dorszewska, "Molecular factors in migraine," *Oncotarget*, vol. 7, no. 31, p. 50708, 2016.

[136] S. M. Neuner, L. A. Wilmott, B. R. Hoffmann, K. Mozhui, and C. C. Kaczorowski, "Hippocampal proteomics defines pathways associated with memory decline and resilience in normal aging and alzheimer's disease mouse models," *Behavioural brain research*, vol. 322, pp. 288–298, 2017.

[137] J. Piñero, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, and L. I. Furlong, "The disgenet knowledge platform for disease genomics: 2019 update," *Nucleic acids research*, vol. 48, no. D1, pp. D845–D855, 2020.

[138] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou *et al.*, "String v10: protein–protein interaction networks, integrated over the tree of life," *Nucleic acids research*, vol. 43, no. D1, pp. D447–D452, 2015.

[139] M. Solarski, H. Wang, H. Wille, and G. Schmitt-Ulms, "Somatostatin in alzheimer's disease: A new role for an old player," *Prion*, vol. 12, no. 1, pp. 1–8, 2018.

[140] E. Burgos-Ramos, A. Hervás-Aguilar, D. Aguado-Llera, L. Puebla-Jiménez, A. Hernández-Pinto, V. Barrios, and E. Arilla-Ferreiro, "Somatostatin and alzheimer's disease," *Molecular and cellular endocrinology*, vol. 286, no. 1-2, pp. 104–111, 2008.

[141] Å. Sandelius, E. Portelius, Å. Källén, H. Zetterberg, U. Rot, B. Olsson, J. B. Toledo, L. M. Shaw, V. M. Lee, D. J. Irwin *et al.*, "Elevated csf gap-43 is alzheimer's disease specific and associated with tau and amyloid pathology," *Alzheimer's & Dementia*, vol. 15, no. 1, pp. 55–64, 2019.

[142] S. De la Monte, S.-C. Ng, and D. W. Hsu, "Aberrant gap-43 gene expression in alzheimer's disease." *The American journal of pathology*, vol. 147, no. 4, p. 934, 1995.

[143] J. Wang, J. Gong, L. Li, Y. Chen, L. Liu, H. Gu, X. Luo, F. Hou, J. Zhang, and R. Song, "Neurexin gene family variants as risk factors for autism spectrum disorder," *Autism Research*, vol. 11, no. 1, pp. 37–43, 2018.

[144] A. Martinez-Mir, A. González-Pérez, J. Gayán, C. Antúnez, J. Marín, M. Boada, J. M. Lopez-Arrieta, E. Fernández, R. Ramírez-Lorca, M. E. Sáez *et al.*, "Genetic study of neurexin and neuroligin genes in alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 35, no. 2, pp. 403–412, 2013.

[145] P. Hollingworth, R. Sweet, R. Sims, D. Harold, G. Russo, R. Abraham, A. Stretton, N. Jones, A. Gerrish, J. Chapman *et al.*, "Genome-wide association study of alzheimer's disease with psychotic symptoms," *Molecular psychiatry*, vol. 17, no. 12, pp. 1316–1327, 2012.

[146] H. Fenton, P. Finch, J. Rubin, J. Rosenberg, W. Taylor, V. Kuo-Leblanc, M. Rodriguez-Wolf, A. Baird, H. Schipper, and E. Stopa, "Hepatocyte growth factor (hgf/sf) in alzheimer's disease," *Brain research*, vol. 779, no. 1-2, pp. 262–270, 1998.

[147] N. C. Inestrosa and L. Varela-Nallar, "Wnt signaling in the nervous system and in alzheimer's disease," *Journal of molecular cell biology*, vol. 6, no. 1, pp. 64–74, 2014.

[148] C. T. Baldwin, C. F. Hoth, R. A. Macina, and A. Milunsky, "Mutations in pax3 that cause waardenburg syndrome type i: ten new mutations and review of the literature," *American journal of medical genetics*, vol. 58, no. 2, pp. 115–122, 1995.

[149] X. Wang, O. L. Lopez, R. A. Sweet, J. T. Becker, S. T. DeKosky, M. M. Barmada, F. Y. Demirci, and M. I. Kamboh, "Genetic determinants of disease progression in alzheimer's disease," *Journal of Alzheimer's disease*, vol. 43, no. 2, pp. 649–655, 2015.

[150] F. H. Duits, G. Brinkmalm, C. E. Teunissen, A. Brinkmalm, P. Scheltens, W. M. Van der Flier, H. Zetterberg, and K. Blennow, "Synaptic proteins in csf as potential novel biomarkers for prognosis in prodromal alzheimer's disease," *Alzheimer's research & therapy*, vol. 10, no. 1, pp. 1–9, 2018.

[151] K. Leng, E. Li, R. Eser, A. Piergies, R. Sit, M. Tan, N. Neff, S. H. Li, R. D. Rodriguez, C. K. Suemoto *et al.*, "Molecular characterization of selectively vulnerable neurons in alzheimer's disease," *bioRxiv*, 2020.

[152] D. G. Bosch, F. N. Boonstra, C. Gonzaga-Jauregui, M. Xu, J. De Ligt, S. Jhangiani, W. Wiszniewski, D. M. Muzny, H. G. Yntema, R. Pfundt *et al.*, "Nr2f1 mutations cause optic atrophy with intellectual disability," *The American Journal of Human Genetics*, vol. 94, no. 2, pp. 303–309, 2014.

[153] W. Machado-Silva, A. C. Tonet-Furioso, L. Gomes, C. Córdova, C. F. Moraes, and O. T. Nóbrega, "The rs4430796 snp of the hnf1β gene associates with type 2 diabetes in older adults," *Revista da Associação Médica Brasileira*, vol. 64, no. 7, pp. 586–589, 2018.

[154] E. Porcellini, I. Carbone, P. L. Martelli, M. Ianni, R. Casadio, A. Pession, and F. Licastro, "Haplotype of single nucleotide polymorphisms in exon 6 of the mzf-1 gene and alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 34, no. 2, pp. 439–447, 2013.

[155] D. Siedlecki-Wullich, J. Català-Solsona, C. Fábregas, I. Hernández, J. Clarimon, A. Lleó, M. Boada, C. A. Saura, J. Rodríguez-Álvarez, and A. J. Miñano-Molina, "Altered micrornas related to synaptic function as potential plasma biomarkers for alzheimer's disease," *Alzheimer's research & therapy*, vol. 11, no. 1, p. 46, 2019.

[156] Z. Lv, L. Hu, Y. Yang, K. Zhang, Z. Sun, J. Zhang, L. Zhang, and Y. Hao, "Comparative study of microrna profiling in one chinese family with psen1 g378e mutation," *Metabolic brain disease*, vol. 33, no. 5, pp. 1711–1720, 2018.

[157] E. Maffioletti, A. Cattaneo, G. Rosso, G. Maina, C. Maj, M. Gennarelli, D. Tardito, and L. Bocchio-Chiavetto, "Peripheral whole blood microrna alterations in major depression and bipolar disorder," *Journal of affective disorders*, vol. 200, pp. 250–258, 2016.

[158] R. M. El-Nashar, N. T. A. Ghani, and S. M. Hassan, "Construction and performance characteristics of new ion selective electrodes based on carbon nanotubes for determination of meclofenoxate hydrochloride," *Analytica chimica acta*, vol. 730, pp. 99–111, 2012.

[159] S.-S. Xie, X. Wang, N. Jiang, W. Yu, K. D. Wang, J.-S. Lan, Z.-R. Li, and L.-Y. Kong, "Multi-target tacrine-coumarin hybrids: cholinesterase and monoamine oxidase b inhibition properties against alzheimer's disease," *European Journal of Medicinal Chemistry*, vol. 95, pp. 153–165, 2015.

[160] E. G. Stopa, K. Q. Tanis, M. C. Miller, E. V. Nikonova, A. A. Podtelezhnikov, E. M. Finney, D. J. Stone, L. M. Camargo, L. Parker, A. Verma *et al.*, "Comparative transcriptomics of choroid plexus in alzheimer's disease, frontotemporal dementia and huntington's disease: implications for csf homeostasis," *Fluids and Barriers of the CNS*, vol. 15, no. 1, p. 18, 2018.

[161] M. Narayanan, J. L. Huynh, K. Wang, X. Yang, S. Yoo, J. McElwee, B. Zhang, C. Zhang, J. R. Lamb, T. Xie *et al.*, "Common dysregulation network in the human prefrontal cortex underlies two neurodegenerative diseases," *Molecular systems biology*, vol. 10, no. 7, p. 743, 2014.

[162] N. C. Berchtold, D. H. Cribbs, P. D. Coleman, J. Rogers, E. Head, R. Kim, T. Beach, C. Miller, J. Troncoso, J. Q. Trojanowski *et al.*, "Gene expression changes in the course of normal brain aging are sexually dimorphic," *Proceedings of the National Academy of Sciences*, vol. 105, no. 40, pp. 15 605–15 610, 2008.

[163] C. Williams, R. M. Shai, Y. Wu, Y.-H. Hsu, T. Sitzer, B. Spann, C. McCleary, Y. Mo, and C. A. Miller, "Transcriptome analysis of synaptoneurosomes identifies neuroplasticity genes overexpressed in incipient alzheimer's disease," *PloS one*, vol. 4, no. 3, p. e4936, 2009.

[164] W. S. Liang, T. Dunckley, T. G. Beach, A. Grover, D. Mastroeni, D. G. Walker, R. J. Caselli, W. A. Kukull, D. McKeel, J. C. Morris *et al.*, "Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain," *Physiological genomics*, vol. 28, no. 3, pp. 311–322, 2007.

[165] O. C. Maes, H. M. Schipper, H. M. Chertkow, and E. Wang, "Methodology for discovery of alzheimer's disease blood-based biomarkers," *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, vol. 64, no. 6, pp. 636–645, 2009.

[166] O. C. Maes, S. Xu, B. Yu, H. M. Chertkow, E. Wang, and H. M. Schipper, "Transcriptional profiling of alzheimer blood mononuclear cells by microarray," *Neurobiology of aging*, vol. 28, no. 12, pp. 1795–1809, 2007.

[167] T. G. Lesnick, S. Papapetropoulos, D. C. Mash, J. Ffrench-Mullen, L. Shehadeh, M. De Andrade, J. R. Henley, W. A. Rocca, J. E. Ahlskog, and D. M. Maraganore, "A genomic pathway approach to a complex disease: axon guidance and parkinson disease," *PLoS genetics*, vol. 3, no. 6, p. e98, 2007.

[168] C. R. Scherzer, A. C. Eklund, L. J. Morse, Z. Liao, J. J. Locascio, D. Fefer, M. A. Schwarzschild, M. G. Schlossmacher, M. A. Hauser, J. M. Vance *et al.*, "Molecular markers of early parkinson's disease based on gene expression in blood," *Proceedings of the National Academy of Sciences*, vol. 104, no. 3, pp. 955–960, 2007.

[169] A. A. Dijkstra, A. Ingrassia, R. X. de Menezes, R. E. van Kesteren, A. J. Rozemuller, P. Heutink, and W. D. van de Berg, "Evidence for immune response, axonal dysfunction and reduced endocytosis in the substantia nigra in early stage parkinson's disease," *PloS one*, vol. 10, no. 6, p. e0128651, 2015.

[170] A. K. Alieva, M. I. Shadrina, E. V. Filatova, A. V. Karabanov, S. N. Illarioshkin, S. A. Limborska, and P. A. Slominsky, "Involvement of endocytosis and alternative splicing in the formation of the pathological process in the early stages of parkinson's disease," *BioMed research international*, vol. 2014, 2014.

[171] X. Xu, Y. Tay, B. Sim, S.-I. Yoon, Y. Huang, J. Ooi, K. H. Utami, A. Ziaei, B. Ng, C. Radulescu *et al.*, "Reversal of phenotypic abnormalities by crispr/cas9-mediated gene correction in huntington disease patient-derived induced pluripotent stem cells," *Stem cell reports*, vol. 8, no. 3, pp. 619–633, 2017.

[172] F. Dangond, D. Hwang, S. Camelo, P. Pasinelli, M. P. Frosch, G. Stephanopoulos, G. Stephanopoulos, R. H. Brown Jr, and S. R. Gullans, "Molecular signature of late-stage human als revealed by expression profiling of postmortem spinal cord gray matter," *Physiological genomics*, vol. 16, no. 2, pp. 229–239, 2004.

[173] G. Morello, M. Guarnaccia, A. G. Spampinato, V. La Cognata, V. D'Agata, and S. Cavallaro, "Copy number variations in amyotrophic lateral sclerosis: Piecing the mosaic tiles together through a systems biology approach," *Molecular neurobiology*, vol. 55, no. 2, pp. 1299–1322, 2018.

[174] S. Corti, M. Nizzardo, C. Simone, M. Falcone, M. Nardini, D. Ronchi, C. Donadoni, S. Salani, G. Riboldi, F. Magri *et al.*, "Genetic correction of human induced pluripotent stem cells from patients with spinal muscular atrophy," *Science translational medicine*, vol. 4, no. 165, pp. 165ra162–165ra162, 2012.

[175] A. S. Chen-Plotkin, F. Geser, J. B. Plotkin, C. M. Clark, L. K. Kwong, W. Yuan, M. Grossman, V. M. Van Deerlin, J. Q. Trojanowski, and V. M.-Y. Lee, "Variations in the progranulin gene affect global gene expression in frontotemporal lobar degeneration," *Human molecular genetics*, vol. 17, no. 10, pp. 1349–1362, 2008.

[176] S. Almeida, Z. Zhang, G. Coppola, W. Mao, K. Futai, A. Karydas, M. D. Geschwind, M. C. Tartaglia, F. Gao, D. Gianni *et al.*, "Induced pluripotent stem cell models of progranulin-deficient frontotemporal dementia uncover specific reversible neuronal defects," *Cell reports*, vol. 2, no. 4, pp. 789–798, 2012.

[177] A. Kemppinen, J. Kaprio, A. Palotie, and J. Saarela, "Systematic review of genome-wide expression studies in multiple sclerosis," *BMJ open*, vol. 1, no. 1, p. e000053, 2011.

[178] W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke *et al.*, "Orchestrating high-throughput genomic analysis with bioconductor," *Nature methods*, vol. 12, no. 2, p. 115, 2015.

[179] S. Davis and P. S. Meltzer, "Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor," *Bioinformatics*, vol. 23, no. 14, pp. 1846–1847, 2007.

[180] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for rna-sequencing and microarray studies," *Nucleic acids research*, vol. 43, no. 7, pp. e47–e47, 2015.

[181] R. Gentleman, V. Carey, W. Huber, and F. Hahne, "Genefilter: methods for filtering genes from high-throughput experiments," *R package version*, vol. 1, no. 1, 2015.

[182] A. Alexa and J. Rahnenfuhrer, "topgo: enrichment analysis for gene ontology," *R package version*, vol. 2, no. 0, p. 2010, 2010.

[183] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang, "Gosemsim: an r package for measuring semantic similarity among go terms and gene products," *Bioinformatics*, vol. 26, no. 7, pp. 976–978, 2010.

[184] G. Yu, L.-G. Wang, G.-R. Yan, and Q.-Y. He, "Dose: an r/bioconductor package for disease ontology semantic and enrichment analysis," *Bioinformatics*, vol. 31, no. 4, pp. 608–609, 2014.

[185] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, "clusterprofiler: an r package for comparing biological themes among gene clusters," *Omics: a journal of integrative biology*, vol. 16, no. 5, pp. 284–287, 2012.

[186] D. Warde-Farley, S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes *et al.*, "The genemania prediction server: biological network integration for gene prioritization and predicting gene function," *Nucleic acids research*, vol. 38, no. suppl_2, pp. W214–W220, 2010.

[187] M. H. Ullman-Cullere and J. P. Mathew, "Emerging landscape of genomics in the electronic health record for personalized medicine," *Human mutation*, vol. 32, no. 5, pp. 512–516, 2011.

[188] K. Nakamura, M. Takeda, T. Tanaka, J. Tanaka, Y. Kato, K. Nishinuma, and T. Nishimura, "Glial fibrillary acidic protein stimulates proliferation and immunoglobulin synthesis of lymphocytes from alzheimer's disease patients." *Methods and findings in experimental and clinical pharmacology*, vol. 14, no. 2, pp. 141–149, 1992.

[189] J. Rafałowska and A. Podlecka, "Does the pathological factor in amyotrophic lateral sclerosis (als) damage also astrocytes?" *Folia neuropathologica*, vol. 36, no. 2, pp. 87–93, 1998.

[190] N. Norgren, P. Sundström, A. Svenningsson, L. Rosengren, T. Stigbrand, and M. Gunnarsson, "Neurofilament and glial fibrillary acidic protein in multiple sclerosis," *Neurology*, vol. 63, no. 9, pp. 1586–1590, 2004.

[191] J. M. Shoffner, "Oxidative phosphorylation defects and alzheimer's disease," *Neurogenetics*, vol. 1, no. 1, pp. 13–19, 1997.

[192] J. M. Shoffner, R. L. Watts, J. L. Juncos, A. Torroni, and D. C. Wallace, "Mitochondrial oxidative phosphorylation defects in parkinson's disease," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 30, no. 3, pp. 332–339, 1991.

[193] M. Martınez, E. Fernández, A. Frank, C. Guaza, M. de la Fuente, and A. Hernanz, "Increased cerebrospinal fluid camp levels in alzheimer's disease," *Brain research*, vol. 846, no. 2, pp. 265–267, 1999.

[194] Y. Kong, X. Liang, L. Liu, D. Zhang, C. Wan, Z. Gan, and L. Yuan, "High throughput sequencing identifies micrornas mediating α-synuclein toxicity by targeting neuroactive-ligand receptor interaction pathway in early stage of drosophila parkinson's disease model," *PLoS One*, vol. 10, no. 9, p. e0137432, 2015.

[195] J. Milovanovic, A. Arsenijevic, B. Stojanovic, T. Kanjevac, D. Arsenijevic, G. Radosavljevic, M. Milovanovic, and N. Arsenijevic, "Interleukin-17 in chronic inflammatory neurological diseases," *Frontiers in Immunology*, vol. 11, p. 947, 2020.

[196] G. Ellrichmann, C. Reick, C. Saft, and R. A. Linker, "The role of the immune system in huntington's disease," *Clinical and Developmental Immunology*, vol. 2013, 2013.

[197] M. O. Klein, D. S. Battagello, A. R. Cardoso, D. N. Hauser, J. C. Bittencourt, and R. G. Correa, "Dopamine: functions, signaling, and association with neurological diseases," *Cellular and molecular neurobiology*, vol. 39, no. 1, pp. 31–59, 2019.

[198] T. Nakase and C. C. Naus, "Gap junctions and neurological disorders of the central nervous system," *Biochimica Et Biophysica Acta (BBA)-Biomembranes*, vol. 1662, no. 1-2, pp. 149–158, 2004.

[199] M. S. Satu, M. I. Khan, M. R. Rahman, K. C. Howlader, S. Roy, S. S. Roy, J. M. Quinn, and M. A. Moni, "Diseasome and comorbidities complexities of sars-cov-2 infection with common malignant diseases," *Briefings in Bioinformatics*, 2021.

[200] A. M. Griesinger, D. K. Birks, A. M. Donson, V. Amani, L. M. Hoffman, A. Waziri, M. Wang, M. H. Handler, and N. K. Foreman, "Characterization of distinct immunophenotypes across pediatric brain tumor types," *The Journal of Immunology*, vol. 191, no. 9, pp. 4880–4888, 2013.

[201] K. Chen, H. Xu, Y. Lei, P. Lio, Y. Li, H. Guo, and M. Ali Moni, "Integration and interplay of machine learning and bioinformatics approach to identify genetic interaction related to ovarian cancer chemoresistance," *Briefings in bioinformatics*, 2021.

[202] J. Ivan, D. Agustriawan, S. S. H. Wibowo, A. A. Parikesit, and R. Nurdiansyah, "Iqgap and hspb8: potent biomarkers in low grade gliomas," in *Journal of Physics: Conference Series*, vol. 1192, no. 1. IOP Publishing, 2019, p. 012056.

[203] Q. Shi, S. Bao, L. Song, Q. Wu, D. Bigner, A. Hjelmeland, and J. Rich, "Targeting sparc expression decreases glioma cellular survival and invasion associated with reduced activities of fak and ilk kinases," *Oncogene*, vol. 26, no. 28, pp. 4084–4094, 2007.

PUBLICATIONS

i Chowdhury, Utpala Nanda, M. Babul Islam, Shamim Ahmad, and Mohammad Ali Moni. "Network-based identification of genetic factors in ageing, lifestyle and type 2 diabetes that influence to the progression of Alzheimer's disease." Informatics in Medicine Unlocked 19 (2020): 100309. DOI:10.1016/j.imu.2020.100309

ii Chowdhury, Utpala Nanda, M. Babul Islam, Shamim Ahmad, and Mohammad Ali Moni. "Systems biology and bioinformatics approach to identify gene signatures, pathways and therapeutic targets of Alzheimer's disease." Informatics in Medicine Unlocked 21 (2020): 100439. DOI:10.1016/j.imu.2020.100439

iii Chowdhury, Utpala Nanda, Shamim Ahmad, M. Babul Islam, Salem A. Alyami, Julian MW Quinn, Valsamma Eapen, and Mohammad Ali Moni. "System biology and bioinformatics pipeline to identify comorbidities risk association: Neurodegenerative disorder case study." PloS one 16, no. 5 (2021): e0250660. DOI: 10.1371/journal.pone.0250660

Informatics in Medicine Unlocked 19 (2020) 100309

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: http://www.elsevier.com/locate/imu

Network-based identification of genetic factors in ageing, lifestyle and type 2 diabetes that influence to the progression of Alzheimer's disease

Utpala Nanda Chowdhury [a], M. Babul Islam [b], Shamim Ahmad [a], Mohammad Ali Moni, PhD, Research Fellow (Moni)[c,d,*]

[a] Dept. of Computer Science and Engineering, University of Rajshahi, Rajshahi, Bangladesh
[b] Department of Electrical and Electronic Engineering, University of Rajshahi, Bangladesh
[c] Pabna University of Science and Technology, Pabna, Bangladesh
[d] WHO Collaborating Centre for eHealth, School of Public Health and Community Medicine, Faculty of Medicine, UNSW Sydney, Australia

ARTICLE INFO

Keywords:
Network-based
Topological
Alzheimer's
Type 2 diabetes
Ageing
Lifestyle

ABSTRACT

Alzheimer's disease (AD) is a progressive neurodegenerative disease, the causes of which are poorly understood, although a number of strong risk factors for AD are known. Understanding how these risk factors affect cell pathways that are altered in AD could identify important causal pathways that could be targeted by therapeutics. We thus pro- posed network-based quantitative frameworks to investigate these risk factor-AD relation- ships. We analyzed gene expression microarray datasets from tissues affected by AD as well as by ageing, high alcohol consumption, type II diabetes (T2D), high dietary fat, obesity, high dietary red meat, sedentary lifestyle, and smoking. These datasets derived from studies that compared tissues affected by these factors with control tissues (not exposed to these factors) to identify differentially expressed genes (DEGs) specific to the risk factors. We employed these to develop gene association and diseasome networks based on neighborhood-based benchmarking and multilayer network topology. We identified 484 DEGs from AD brain tissue, of which 27 were also seen in the smoking DEGs gene set. AD data also showed 21 DEGs in common with T2D, and 12 with sedentary lifestyle datasets. AD shared less than ten DEGs with the other factors, but 3 (HLA-DRB4, IGH and IGHA2) were commonly up-regulated among the AD, T2D and high alcohol consumption datasets. IGHD and IGHG1 were up-regulated among AD, T2D, alcohol and sedentary lifestyle datasets. Protein-protein interaction networks identified 10 hub genes: CREBBP, PRKCB, ITGB1, GAD1, GNB5, PPP3CA, CABP1, SMARCA4, SNAP25 and GRIA1. Ontological and pathway analyses identified significant gene ontology and molecular pathways that could enhance our understanding of the mechanisms of AD progression by suggesting new therapeutic approaches to affect the development of AD. We verified genes from ontological and pathway analyses with gold benchmark gene-disease associations databases including Online Mendelian Inheritance in Man (OMIM) and dbGaP. This supports our identification of disease associations for the putative AD target genes. These outcomes provide further evidence that network-based approaches can generate new insights into AD progression.

1. Introduction

Alzheimer's disease (AD) is the most common form of dementia and is characterized by gradual degeneration in memory, cognitive processes, language use, and learning capacity [1]. Initial indications begin with a reduced ability to retain recent memories, but with progression all cognitive functions are inevitably affected, resulting in complete dependency for basic daily activities and greatly increased risk of premature death. At present AD accounts for 60% to 80% of all dementia cases and is estimated to affect over 24 million people worldwide. In the United States, 93,541 deaths resulting from AD were officially recorded in 2014, which is ranked sixth among all causes of death in the United States and fifth among all causes of death after 65 years of age [2,3]. The premature death rate for AD sufferers increased by 89% in the five years up to 2010, whereas death rates associated with other major morbidities such as cardiac disease, stroke, breast and prostate cancer, and AIDS all declined in that time frame [4–7].

The pathogenesis of the AD is not clearly understood, but it is clear

* Corresponding author. Pabna University of Science & Technology University, Bangladesh.
E-mail address: m.moni@unsw.edu.au (M.A. Moni).

https://doi.org/10.1016/j.imu.2020.100309
Received 4 December 2019; Received in revised form 10 February 2020; Accepted 2 March 2020
Available online 5 March 2020

Figure 5.1: Publication 1

Informatics in Medicine Unlocked 21 (2020) 100439

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: http://www.elsevier.com/locate/imu

ELSEVIER

Systems biology and bioinformatics approach to identify gene signatures, pathways and therapeutic targets of Alzheimer's disease

Utpala Nanda Chowdhury [a], M. Babul Islam [b], Shamim Ahmad [a], Mohammad Ali Moni, PhD [c,d,*]

[a] *Department of Computer Science and Engineering, University of Rajshahi, Rajshahi, Bangladesh*
[b] *Department of Electrical and Electronic Engineering, University of Rajshahi, Bangladesh*
[c] *Department of Computer Science and Engineering, Pabna University of Science and Technology, Pabna, Bangladesh*
[d] *WHO Collaborating Centre on eHealth, School of Public Health and Community Medicine, Faculty of Medicine, University of New South Wales, Sydney, Australia*

ARTICLE INFO

Keywords:
Alzheimer's disease (AD)
cis-eQTL
GWAS
Brain
Blood

ABSTRACT

Alzheimer's disease (AD) develops relentlessly in affected individuals and its occurrence is increasing. A clinical test to diagnose early-stage AD could be an important means of enabling interventions to slow its progression. However, available neuroimaging and cerebrospinal fluid-based diagnoses are very costly. Therefore, detecting AD from blood transcripts that mirror the expression of brain transcripts in the AD could improve the diagnosis. To achieve this goal, we employed a transcriptional analysis of affected tissues and integrated them with cis-eQTL data. In this study, we analyzed microarray gene expression data of brain and blood cells from AD patients and control individuals. Differentially expressed genes (DEGs) common to both brain tissue and blood cells were identified. Potential common genes and molecular pathways were identified using overlapping DEGs through the pathway and gene ontology enrichment analysis. We identified 18 significantly dysregulated genes shared by both brain and blood cells in AD affected individuals. We validated these candidates as disease-associated genes using gold-standard benchmarking databases (gene SNP-disease linkage). Significant molecular pathway and gene ontology indicating AD progression were identified. This study also identified regulatory factors, including transcription factors (TFs), microRNAs and candidate drugs. In sum, we identified new putative links between pathological processes in brain tissue and blood cells in AD that may allow assessment of AD status using blood samples. Thus, our formulated methodologies demonstrate the power of gene and gene expression analysis for brain-related pathologies transcriptomics, cis-eQTL, and epigenetics data from brain and blood cells.

1. Introduction

Alzheimer's disease (AD) is a chronic, progressive and neurodegenerative disease symptomized by gradual degeneration in memory, thinking, language, and learning capacity, leading to full-blown dementia [1,2], resulting in complete dependency for basic daily activities, and fostering premature death [3,4]. It is an irremediable disease that is responsible for 60–80% of all dementia cases and affecting over 24 million people worldwide. In the United States alone, 93,541 deaths from AD were officially recorded in 2014 making AD the sixth-ranked cause of death. In addition, it is notable that deaths attributable to AD have increased by 89% within five years till 2010, whereas death rates of other major diseases like cardiac disease, stroke, breast and prostate cancer, and AIDS have declined in the same time frame. Currently, in every 66 s one new case of the AD is developed, which is estimated to become 2 times faster by 2050, triggering nearly 1 million new cases per year [5,6].

Numerous research approaches throughout the last century enriched our understanding of the factors that influence the AD progression [7–9] and reported a significant number of molecular signatures and therapeutic agents [10,11]. Still, they could not successfully find any cure or disease-modifying drug to treat AD effectively. Most of the AD-related pathology occurs in the brain [12,13]. Thus, it is immensely challenging to study brain tissue because these tissue samples have to be collected post-mortem with a high degree of cellular heterogeneity. Hence, there is no definite early pre-mortem diagnosis for AD aside from cognitive assessments and the use of brain imaging methods that reveal significant degeneration. Symptoms of degradation in cognition due to AD emerge only after significant, unchangeable neural degeneration has occurred, hence there is a need for a simple means to detect AD early,

* Corresponding author. Department of Computer Science and Engineering, Pabna University of Science and Technology, Pabna, Bangladesh.
E-mail addresses: m.moni@unsw.edu.au, moni@pust.ac.bd (M.A. Moni).

https://doi.org/10.1016/j.imu.2020.100439
Received 15 August 2020; Received in revised form 21 September 2020; Accepted 27 September 2020
Available online 10 October 2020

Figure 5.2: Publication 2

PLOS ONE

RESEARCH ARTICLE

System biology and bioinformatics pipeline to identify comorbidities risk association: Neurodegenerative disorder case study

Utpala Nanda Chowdhury[1], Shamim Ahmad[1], M. Babul Islam[2], Salem A. Alyami[3], Julian M. W. Quinn[4], Valsamma Eapen[5], Mohammad Ali Moni[4,5,6]*

1 Department of Computer Science and Engineering, University of Rajshahi, Rajshahi, Bangladesh, **2** Department of Electrical and Electronic Engineering, University of Rajshahi, Rajshahi, Bangladesh, **3** Department of Mathematics and Statistics, Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia, **4** Healthy Ageing Theme, Garvan Institute of Medical Research, Darlinghurst, NSW, Australia, **5** School of Psychiatry, Faculty of Medicine, University of New South Wales, Sydney, Australia, **6** WHO Collaborating Centre on eHealth, School of Public Health and Community Medicine, Faculty of Medicine, UNSW Sydney, Sydney, Australia

* m.moni@unsw.edu.au

Check for updates

Citation: Chowdhury UN, Ahmad S, Islam MB, Alyami SA, Quinn JMW, Eapen V, et al. (2021) System biology and bioinformatics pipeline to identify comorbidities risk association: Neurodegenerative disorder case study. PLoS ONE 16(5): e0250660. https://doi.org/10.1371/journal.pone.0250660

Editor: Patrick Lewis, University College London Institute of Neurology, UNITED KINGDOM

Received: December 3, 2020

Accepted: April 12, 2021

Published: May 6, 2021

Data Availability Statement: All data are publicly available at https://www.ncbi.nlm.nih.gov/geo/ with accession numbers: GSE110226, GSE12685, GSE1297, GSE13162, GSE21942, GSE27206, GSE33000, GSE40378, GSE4226, GSE4229, GSE48350, GSE49036, GSE5281, GSE54536, GSE6613, GSE68605, GSE7621, GSE833, GSE93767.

Abstract

Alzheimer's disease (AD) is the commonest progressive neurodegenerative condition in humans, and is currently incurable. A wide spectrum of comorbidities, including other neurodegenerative diseases, are frequently associated with AD. How AD interacts with those comorbidities can be examined by analysing gene expression patterns in affected tissues using bioinformatics tools. We surveyed public data repositories for available gene expression data on tissue from AD subjects and from people affected by neurodegenerative diseases that are often found as comorbidities with AD. We then utilized large set of gene expression data, cell-related data and other public resources through an analytical process to identify functional disease links. This process incorporated gene set enrichment analysis and utilized semantic similarity to give proximity measures. We identified genes with abnormal expressions that were common to AD and its comorbidities, as well as shared gene ontology terms and molecular pathways. Our methodological pipeline was implemented in the R platform as an open-source package and available at the following link: https://github.com/unchowdhury/AD_comorbidity. The pipeline was thus able to identify factors and pathways that may constitute functional links between AD and these common comorbidities by which they affect each others development and progression. This pipeline can also be useful to identify key pathological factors and therapeutic targets for other diseases and disease interactions.

Introduction

Alzheimer's disease (AD) is the most frequent neurodegenerative disease (NDD) which is considered to be the current primary cause of dementia, causing most of all dementia cases (60%

Figure 5.3: Publication 3