T.R.

GEBZE TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

NETWORK-BASED ANALYSIS OF COGNITIVE IMPAIRMENT AND MEMORY DEFICITS FROM TRANSCRIPTOME DATA

ELİF EMANETCİ

A THESIS SUBMITTED FOR THE DEGREE OF

MASTER OF SCIENCE DEPARTMENT OF BIOENGINEERING

BIOINFORMATICS AND SYSTEMS BIOLOGY PROGRAMME

GEBZE

2019

T.R.

GEBZE TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

NETWORK-BASED ANALYSIS OF COGNITIVE IMPAIRMENT AND MEMORY DEFICITS FROM TRANSCRIPTOME DATA

ELİF EMANETCİ

A THESIS SUBMITTED FOR THE DEGREE OF MASTER OF SCIENCE DEPARTMENT OF BIOENGINEERING BIOINFORMATICS AND SYSTEMS BIOLOGY PROGRAMME

THESIS SUPERVISOR

ASSOC. PROF. DR. TUNAHAN ÇAKIR

GEBZE

2019

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TRANSKRİPTOM VERİLERİ KULLANILARAK HAFIZA PROBLEMLERİNİN VE BİLİŞSEL BOZUKLUKLARIN HÜCRESEL AĞLARA DAYALI ANALİZİ

ELİF EMANETCİ
YÜKSEK LİSANS TEZİ
BİYOMÜHENDİSLİK ANABİLİM DALI
BİYOİNFORMATİK VE SİSTEM BİYOLOJİSİ PROGRAMI

DANIŞMANI
DOÇ. DR. TUNAHAN ÇAKIR

GEBZE
2019

GEBZE TEKNİK ÜNİVERSİTESİ	**YÜKSEK LİSANS JÜRİ ONAY FORMU**

GTÜ Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 03/07/2019 tarih ve 2019/30 sayılı kararıyla oluşturulan jüri tarafından 08/07/2019 tarihinde tez savunma sınavı yapılan Elif EMANETCİ'nin tez çalışması BİYOMÜHENDİSLİK Anabilim Dalında Biyoinformatik ve Sistem Biyolojisi Programında YÜKSEK LİSANS tezi olarak kabul edilmiştir.

JÜRİ

ÜYE (TEZ DANIŞMANI)	: Doç. Dr. Tunahan ÇAKIR
ÜYE	: Doç. Dr. Fatih Erdoğan SEVİLGEN
ÜYE	: Dr. Öğretim Üyesi Hilal TAYMAZ NİKEREL

ONAY

Gebze Teknik Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun

…/.../…. tarih ve ……/….. sayılı kararı.

SUMMARY

Brain is the most important organ in our body. It controls vital functions and all cognitive mechanisms. Aging is the most crucial factor that affects brain functioning. Numerous experimental studies were conducted in the literature to investigate the effect of aging on learning and memory performance by using model organisms. In those studies, memory tests were applied to the organism, and transcriptome data was obtained from hippocampus region, which is accepted as the learning center of the brain. These studies are limited in terms of elucidating mechanisms because the transcriptome data was not mapped on molecular interaction networks. They only identified differentially expressed genes to associate aging with memory. Subnetwork discovery and Network Inference are useful bioinformatics approaches for more efficient interpretation of omics data, based on molecular interactions between proteins. In this thesis study, transcriptome data of learning and memory related experiments from *Rattus norvegicus* were downloaded from the Gene Expression Omnibus (GEO) database, and computationally analyzed. Two methods were alternatively used for each approach; Bionet and KPM for Subnetwork Discovery approach, and WGCNA and Pearson Correlation for Network Inference approach. The first approach enabled the identification of subnetworks by mapping transcriptome data on protein-protein interaction networks while the second one led to modules consisting of highly correlating genes. Functional analysis was performed on the identified modules and subnetworks. Our study analyzes transcriptome data related to memory and cognitive disorders together with molecular interaction networks for the first time, contributing to the elucidation of molecular mechanisms behind such disorders.

Keywords: Memory and Learning, Transcriptome, Protein Interaction Networks, Correlation Networks, Systems Biology.

ÖZET

Hayati fonksiyonlarımızın kontrolünden tüm bilişsel mekanizmaların çalışmasına kadar birçok görevi olan beyin vücudumuzdaki en önemli organımızdır. Beynin çalışmasını negatif yönde etkileyen birçok etmen vardır ve yaşlanma bunların başında gelir. Yaşlanmanın öğrenme ve hafıza üzerindeki etkisini araştırmak amacıyla literatürde çok sayıda deneysel çalışma yapılmıştır ve bu çalışmalarda model organizma olarak sıçan (*Rattus Norvegicus*) kullanılmıştır. Sıçanlara hafıza testleri uygulanmış, hippocampus bölgesi çıkarılıp transkriptom verisi elde edilmiştir. Transkriptom deneylerinde anlamlı değişen genler hesaplanıp bu genlerin foksiyonel analizleri yapılarak yaşlanma sonucu oluşan bilişsel bozukluklarla ilişkilendirilmiştir. Bu çalışmalar literatür açısından zengin bir veri seti sunsa da, omik verinin moleküler etkileşim ağlarıyla birlikte incelenmesini içermemeleri açısından eksik bir yaklaşım sergilemektedir. Günümüzde çeşitli biyoinformatik yöntemler kullanılıp geliştirilerek büyük veriler işlenip anlamlandırılabilmektedir. Moleküler etkileşimleri de dikkate alan Alt-ağyapı keşfi (*İng. Subnetwork discovery),* ve Ağyapı çıkarımı *(İng. Network Inference*) yöntemleri, yeni nesil omik verilerin daha verimli bir şekilde anlamlandırılması açısından önemli ve kullanışlı biyoinformatik yaklaşımlardır. Bu tez çalışmasında hafıza ve bilişsel bozuklukların mekanizmasını anlamak için tasarlanan deneylerden elde edilen sıçan transkriptom verileri Gene Expression Omnibus veri tabanından indirildi ve iki alternatif alt-ağyapı keşfi yöntemi (BioNet ve KPM) kullanılarak protein etkileşim ağlarına dayalı alt-ağlar oluşturuldu. İkinci yöntem olan Ağyapı çıkarımı için de WGCNA algoritması ve Pearson korelasyonu yöntemleri karşılaştırmalı olarak kullanıldı. Bu yöntemlerle, gen çiftlerine ait gen ekspresyonu verileri arasındaki korelasyon hesaplanarak moleküler etkileşim ağları oluşturuldu, ve bu ağlar modüllere bölündü, fonksiyonel olarak incelendi. Çalışmamız; hafıza ve bilişsel bozukluklar ile ilgili transkriptom verilerinin bahsedilen yöntemlere ilk defa girdi olarak verilmesini sağlamıştır. Ayrıca öngördüğü etkileşim ağlarıyla bu bozuklukların moleküler mekanizmalarının aydınlatılmasına katkı sağlayacaktır.

Anahtar Kelimeler: Hafıza ve Öğrenme, Transkriptom, Korelasyon, Protein Etkileşimi, Sistem Biyolojisi

ACKNOWLEDGEMENTS

I deeply thank my supervisor Assoc. Prof. Tunahan Çakır for his patience and perfect guidance. He always encouraged me and convinced me that I would do good job. I am so glad to study under his supervision during my master study.

I want to thank my office mates for their very nice collaborations and help. Especially, I would like to thank Emel Kökrek and Merve Kutay for their friendship and patience.

I would like to express my endless thanks to my family including my parents. I especially thank my sister, Fatma EMANETCİ, for her patience, help, and for her friendship during the thesis process.

TABLE of CONTENTS

LIST of ABBREVIATIONS and ACRONYMS

Abbreviations and Acronyms		Explanations
21PT	:	21 Day After Training (21 days post training)
5T	:	5 Day After Training (Training complete)
A	:	Aged
AI	:	Aged-Impaired
AU	:	Aged-Unimpaired
BP	:	Biological Processes
CC	:	Cellular Component
FDR	:	False Discovery Rate
GEO	:	Gene Expression Omnibus
GO	:	Gene Ontology
GS	:	Gene Significance
Keg	:	Kegg Database
KPM	:	Key Pathway Miner
MD	:	Middle aged
MF	:	Molecular Function
MM	:	Module Membership
MWM	:	Morris Water Maze
NT	:	Non-Trained
OMT	:	Object Memory Task
PCA	:	Principle Component Analysis
PCC	:	Pearson Correlation Coefficient
PPI	:	Protein-Protein Interaction
Rea	:	Reactome Database
RMA	:	Robust Multichip Average
WGCNA	:	Weighted Gene Correlation Network Analysis
Y	:	Young

LIST of FIGURES

LIST of TABLES

1. INTRODUCTION

Brain is a central part of the central nervous system, and all vertebrates have brain. It is the most important organ in organisms and controls the functions of other organs. Brain has other functions such as control of vital functions, receiving, processing, integrating and interpreting all information that are received through our senses, control of our movements. It is also responsible for feelings and behavior. The most important function of the brain is to control higher cognitive functions: memory, learning, perception, and executive functions [The Editors of Encyclopedia Britannica, 2019]. But like other organs, brain is also affected from the environmental factors and its performance decreases with time, called aging.

In this study, deficits caused due to aging is investigated at molecular level by using transcriptome experiments in literature. There are lots of experimental designs in literature to understand the aging mechanism in brain. Some of these experiments created transcriptome data [Barbulovic-Nad et al., 2006]. Genome-wide omic data helps us to understand aging mechanisms from systems-wide perspective.

There are several bioinformatics methods in literature to analyse the transcriptomic data, and network-based analyses are among the most poweful ones. These methods either discovers subnetworks by mapping transcriptome data on interactome [Batra et al., 2016] or create co-expression networks by using correlation methods [Werhli et al., 2006], [Albert et al., 2007], [Langfelder et al.,2008].

This study aims to identify the effects of aging on learning and cognitive performance of brain. Chapter 2 explains the details of the aging and memory mechanisms based on the literature information and experiments and gives information about Protein Protein Interaction (PPI) networks. Also, Subnetwork Discovery and Network Inference algorithms are explained in detail in Chapter 2. In Chapter 3 and 4, detailed computational analyse of the two transcriptome datasets are given with an emphasis on molecular interactions. Functional analysis of subnetworks and modulesidentified s outputs of computational analyses are also given at the end of the chapters. These results are compared to the literature and discussed in Chapter 5.

2. BACKGROUND INFORMATION

2.1. Memory

Brain is the most important organ in all organisms. It controls all physiological and psychological processes. Brain is responsible for the communication of body parts. It performs this communication via nervous system, and especially via special cell types like neurons and astrocytes. The brain, or specific parts of the brain, like our other organs, loses or decelerates some functions over time. Environmental factors, stress, fear and insomnia can affect brain and cause neurodegenerative diseases, but the time factor alone is sufficient for brain dysfunction [Crowder et al., 2015]. This functional loss is called aging, and it is an inevitable fact in all organisms. Brain aging is not understood well because of its complexity; it affects multiple systems and molecular processes such as lipid metabolism, Calcium balance, inflammatory processes, mitochondrial function etc. [Blalock et al., 2003]. Brain aging affects cognition and memory performance of organisms, generally in negative direction in course of time.

Cognition is defined as the ability of processing information gained from our experience to interpret our life. With time and aging, this ability deteriorates, and some psychological disturbances can occur [Lazarus et al., 1991]. Cognition includes learning, memory, decision making and attention. Memory is, in general definition, a process of taking the information from the environment, processing and understanding it and finally storing it for recalling when needed [Rankin et al., 1990]. An alternative definition of memory is the sum of what we remember. It is the power to recall experiences and previously learned cases, habits and skills. Memory is an important requirement for any organism since it also stores the information of strategic adaptation to changing environment, which is important to survive [Rasch et al., 2013].

2.2. Molecular Mechanisms of Memory

Normal aging, without any physical disturbance, has the most important and unidentified impact on brain. There are several experiments in literature designed to understand the mechanisms of the aging in brain. These experiments use mouse as a model organism because they are easily available, and human and mouse have 90% genes in common [Perlman et al., 2016]. *Rattus norvegicus* is used as a model organism for memory and learning experiments. *Mus musculus* is also a model organism for memory and learning related experimental design, but generally it is used to analyse the effect of drugs on specific diseases. In these experiments, animals are exposed to stress condition and their memory and learning performance are measured. Stress condition is created by seperating organisms from their mother while they were newborn [Suri et al., 2013] or wearing them a stress jacket to restrict their movement [Buechel et al., 2014]. There is another type of experiments in the literature that search for the effect of aging on brain by using behavioral responses of animals to memory tests [Verbitsky et al., 2004].

These experiments create a model for aging brain to understand molecular mechanism of aging and its effects on memory performance by using experimental data from the hippocampus region of brain. Blalock et al [Blalock et al., 2003] created an aging brain model by comparing young and aged rats, and this model shows that the disturbance in Ca+ signaling mechanism negatively affects the neuronal activity in cell. The decrease in the activity has side effects on bio-processing mechanisms and affects extracellular mechanisms. This causes internal myelination, and it triggers demyelination. Microglia, which are immune system cells for brain, supports demyelination processes, and astrocytes, the supporter cells in brain, increase in size because glucose transmission increases from blood vessels to astrocytes to support microglia cells. This unbalanced material transmission changes the normal neuronal activity because of nutrient deficiency in neurons. This deficit reduces energy production is required for signaling mechanisms and causes deficits in memory and learning performance (Figure 2.1) [Blalock et al., 2003].

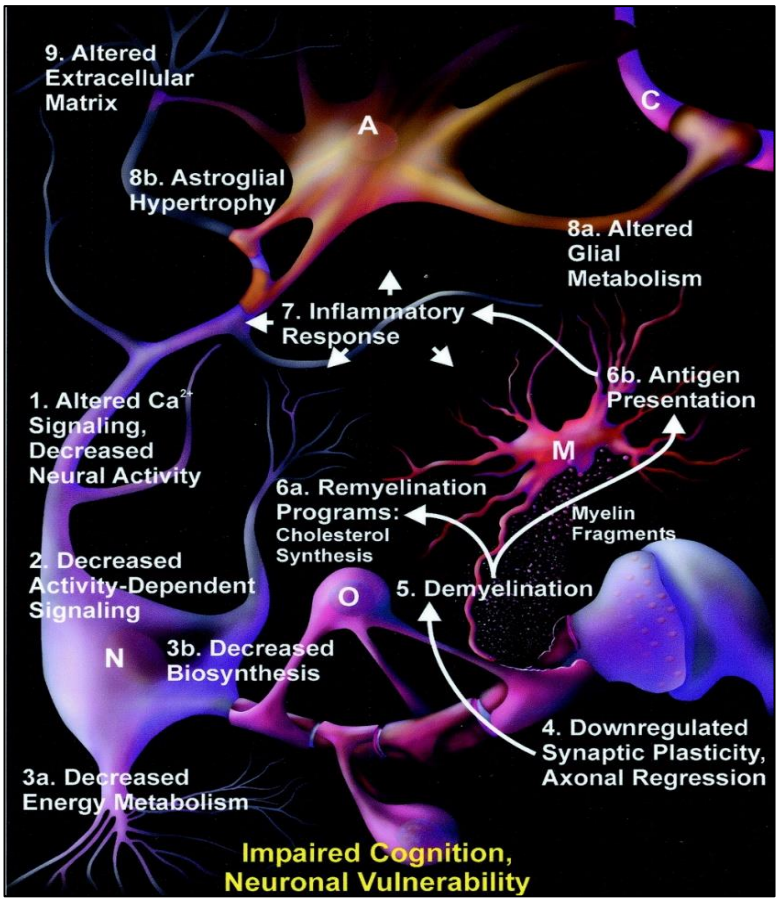

Figure 2.1: Integrative model for brain aging about cognition.

Rowe et al. [Rowe et al., 2007] created another rat-based aging brain model. In this experiment, aged animals were separated into two groups; impaired and unimpaired, based on their Morris Water Maze performances when compared to the young animals. In this experiment, transcriptome data was statistically analysed, and up-down regulated mechanisms were discovered. Based on these mechanisms, relationships were suggested between some molecular processes to explain the effect of aging in brain. These relationships, given in Figure 2.2 [Rowe et al., 2007], constitute their model for aging brain. This model shows that impaired animals have maintanance problems about glucose utilization. Energy providing astrocytic processes were downregulated in impaired animals and this causes downregulation in neuronal pathways. Then, signaling mechanisms and substrate balance are disturbed. Notch signaling mechanisms are sensitive to the injury-like signals and upregulated. Oligodendrocytes are activated, and this triggers myelination processes. Myelination suppresses neuronal pathways or plasticity mechanisms, which can be defined as the ability of brain to regulate or create new connections [Johansson et al., 2012]. If

plasticity mechanisms are disturbed, not only human brain but also brain of other organisms cannot improve and cannot heal itself against any damage This trigger memory deficits (Figure 2.2) [Rowe et al., 2007].

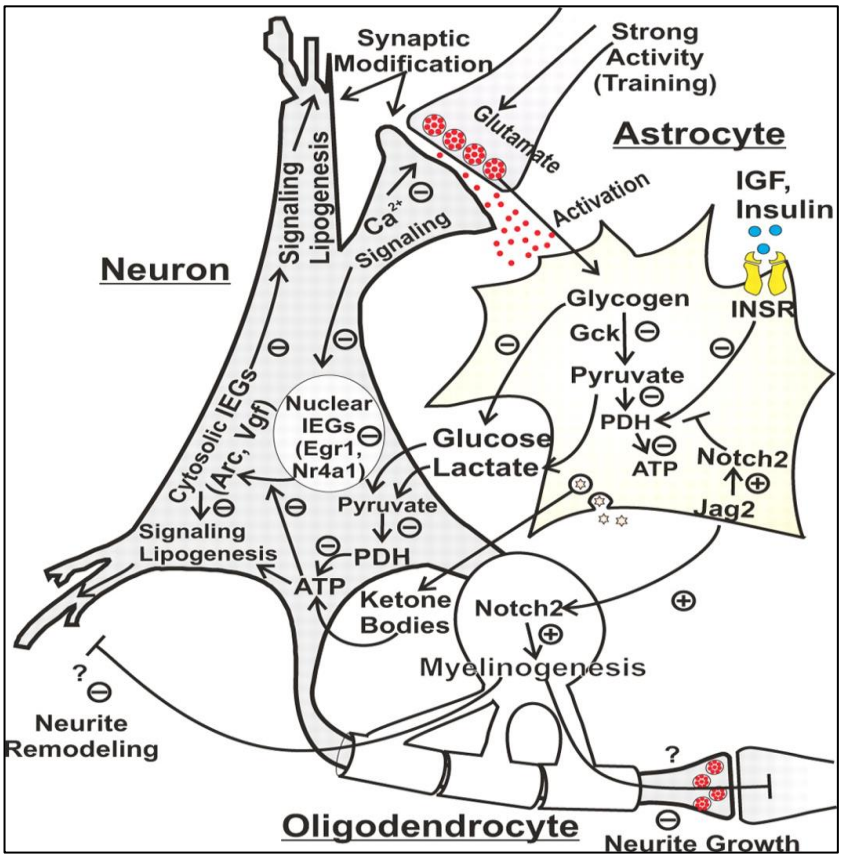

Figure 2.2: Schematic model of alteration in impaired animals in aging brain.

There is some important mechanism that are affected from aging and affects the memory performance such as synaptic plasticity, gliogenesis, neurogenesis, myelination and demyelination, cytokine and signaling mechanisms.

2.2.1. Synaptic Plasticity

Brain stores the information that were received through senses and this information is transmitted from neuron to neuron. This transmission can be done physically or chemically. All neurons have synapses, and they chemically transport the information through synapses. Brain stores information by changing synaptic connection, called synaptic plasticity. In general definition, synaptic plasticity is the

ability of brain to change. The connection between neuron cells is not static but thay are plastic, they can be changed and affected with new signals and stimulants [Heinbockel et al., 2017]. This ability helps brain to understand and store information called memory. The most important and effective damager for synaptic plasticity and memory is time, called aging. Neurons must be strong and have resistance in case of movement and damage. But they also must be plastic to adapt to changing environment. These two opposite phenomena cause dilemma in terms of consuming resources and defining requirements, leading to progressive memory deficit and synaptic plasticity impairment during aging [Baudry et al., 2017].

2.2.2. Gliogenesis and Neurogenesis

Gliogenesis and neurogenesis are related terms. Neurogenesis is about modifying neuronal cells and their connection. Gliogenesis covers production of supporting glia cells, oligodentrocytes and astrocytes and responsible for myelination mechanisms. Neurogenesis also produces neurons from neuronal stem cells, and these neurons are used in Learning and Memory mechanisms. Altered gliogenesis and neurogenesis mechanisms cause neurodegenerative, neuropsychiatric and demyelinating diseases by affecting central nervous system. Time and aging also affects this mechanism and cause cognition deficits. The modulation and restoration of these processes can be used to treat neurological diseases [Rusznák et al., 2016].

Hippocampus is known as the functional and critical part of the brain about learning and memory. Neurogenesis also occurs in this region, and therefore this process is related to memory mechanisms. In hippocampus, the mechanism of production of neuron is triggered and the number of neurons inceases to upregulate synaptic plasticity during learning. Newly produced neurons are sensitive to the signals and they are more strong than older neurons. Gliogenesis produces the supporter cells for neurons, and these cells provide glucose from blood, protect the neurons from the environment and repair the damaged parts of the neurons, create myelin sheet for faster signal transmission. These mechanisms support neurons and help neurogenesis mechanisms to produce new neurons by providing required resources during learning processes [Kaptan et al., 2016].

2.2.3. Myelination and Demyelination

Myelin is a sheet that covers the axons. These sheets are formed by Central Nervous System or Peripheral Nervous system. In central nervous system, glia cells called oligodentrocytes are responsible for the myelination. Myelination or miyelinogenesis is a process of generation of myelin sheets. Myelin sheets protects axons from the environment. With myelin sheet the signal transmission through neurons gets faster since it prevents the electrical current leakage and saves the energy [Salzer et al., 2016]. Any damage or loss of myelin is defined as demyelination. Physical compressions, oxygen loss, metabolic problems and viral infections cause demyelination. Without myelin sheet, the quality of transmitted information may deteriorate, and transmission gets slower [Love at al., 2006]. Demyelination causes Multiple Sclerosis disease, which is one of the neurodegenerative diseases. In demyelinated brain part, the levels of proteins and essential metabolites are decreased compared to the myelinated parts. Decreasing metabolites cause deficits in neurotransmission, axonal transport, and memory formation. Myelination process is important in synaptic plasticity mechanism, this process prevents axons from damage and responsible for transmission of information safely and help for the creation of new connections [Dutta et al., 2013].

2.2.4. Cell-cell Signaling

Brain consists of neuron, nerve cells and supporter cells called glia. All these cells use signaling mechanism to transport information from cell to cell. Neurons use signalling mechanisms to transfer information that are received through our senses. Matter and energy are transported from cell to cell by using internal signalling mechanisms. Intracellular and intercellular signalling mechanisms create signaling networks that have complex structure to transport stimuli and response. These signalling networks are responsible for specific functions and used to treat any deficits they cause [Koseska et al., 2017].

Memory can be defined as storing information that are received through senses and experiences via transforming physical stimulus to electrical signals. During learning procedure new connections are created and existing connections are changed, called synaptic plasticity, via signalling mechanisms. All metabolic events in srain or

other organs use signalling mechanisms to transport matter and energy, to form new connection, to understand, interpret, integrate and store new information and to communicate with each other. Learning and memory mechanisms require these signaling cascades to work efficiently [Mitra et al., 2018].

2.2.5. Immune System and Cytokine Metabolism

Cytokines are proteins that are responsible for cell interactions and communications. Cytokines are known as messenger molecules; they bind to the receptor and evoke biological activity. Cytokines especially have mission in nervous system, they work for inflammatory responses. In case of nerve injury in neurons or any part of nervous system, macrophages or microglia cells are located around cells, and they produce specific growth factors or cytokines that can be used for regeneration of nerve cells [Zhang et al., 2007]. Immune mechanisms are stimulated by environmental or psychological factors, and they regulate the neuronal circuits by remodeling them, they foster neurogenesis via secretion of cytokines. In higher stress or injury conditions, the morphology of glia and other related brain cells are changed, and they secrete proinflammatory cytokines. Inflammatory processes protect cells from the inflammation condition, but they use brain cells to prevent deficits by remodeling them. The production of mocecules- neurotrophins- that are important for plasticity mechanism are decreased with inflammatory processes. Also, any inflammation in brain cells causes neurodegenerative and neuropsychiatric diseases, and these mechanisms are associated with normal aging [Yirmiya et al., 2011].

2.2.6. Circadian Rhythm

Circadian rhythm is a biological phenomenon and defined as 24-hour regular cycle processes in living beings. This process is controlled by hippocampus. Circadian rhythm regulates physical events such as cell cycle, feeding, body temperature, sleep-wake cycle and metabolism. This process also has an effect on memory and learning performance in animals and cognitive performance on human. Circadian rhythm is affected from environmental stimulants such as sunlight or moonlight, behavior, stress, pharmaceuticals. Genetic code is also an important effector [Gersner et al., 2010]. If there are not any disorders in organisms, age can only be an effector on circadian

rhythm. With time, the signalling mechanisms, the number of heathy neurons and performance of matter and energy transport decrease, and circadian rhythm is also affected negatively. This reverberates the physical and mental performance of an organism. Circadian rhythm regulates lots of mechanisms that include learning and memory. It also affects the maze performance of rats during training. The activity of gene expression and transcription factors differs in different time intervals during the day, and this affects the physical performance of an organsm [Antoniadis et al., 2000].

2.3. Transcriptome Data

Transcriptome is defined as the whole content of gene transcripts in a cell and represents the relation between phenotype and information stored and coded in DNA. Transcriptomic datasets are large datasets, they consist complete set of RNA transcripts produced by genome. With the technology, now the characteristics profiles of whole genome expression can be obtained. Microarray, also called genechip or DNAchip, consists of gene-specific DNA fragments, which are attached on a chip with covalent bonds [Barbulovic-Nad et al., 2006]. Gene expression microarrays can give information about the activity of transcriptome in different types of cells and tissues, the behavior of transcriptome in disease and the transcriptome variances in different organisms [Malone at al., 2011]. Microarrays are used in several experimental designs in genetic and molecular biology. Gene Expression Omnibus (GEO) [Clough et al 2016] is a database that stores high-throughput gene expression and other functional genomics data sets. Data can be freely accessed with its title or accession number. Basic statistical analyses (detecting differential expressed genes) can also be performed online through GEO.

There are several experimental designs in literature that consider enviromental factors, stress, drug effects, aging etc., to understand the effect of these factors on memory performance and cognition deficits in brain. Some of these experiments produced transcriptome data, and these data can be used for advanced bioinformatic analyses.

2.4. Protein-Protein Interactions

Complex biological systems and cellular networks have some genotypic or phenotypic relationships. Protein-protein Interactions are examples for such networks, and the corresponding data, called interactome, represents proteins and their relationship with each other. Interactome data are organism specific. These protein interactions are mostly detected from the results of experiments, and some of them are predicted computationally. [Kusonmano et al., 2016].

Protein interactome data can be obtained from different databases. Major protein interactome databases are BioGrid [Stark et al., 2006], Mint [Licata et al., 2012], Intact [Orchard et al., 2013], Uniprot [The UniProt Consortium, 2019] and iRefindex [Razick et al., 2008]. The PPI information in these databases can be downloaded using PSICQUIC (The Proteomics Standard Initiative Common QUery InterfaCe) [Aranda et al., 2011], a database that collects information from nearly 33 databases. This database uses PSI-MI format that expedite collection of data from different resources. PSI-MI includes information on interactions such as gene name, protein name, organism name, source of experiment, first author of paper, publication date.

2.5. Mapping Techniques of Omic data on Interactome

Human genome stores genetic information in nucleotide sequences, and this information can be read by computers to understand genes and their functions. Computer science is not only used for storing and reading genetic information, but it also contributes to methods for the analysis of data using mathematical methods. The combination of computer science and biology created the scientific field called Bioinformatics. Computers use data that are the results of biological experiments, and they use this information to understand the relationships in data to elucidate molecular mechanisms in health and disease. The science of bioinformatics grows exponentially with the developments in data collection techniques [Lesk et al., 2019]. Bioinformatics approach brings a new point of view to the analysis of biological data, and this increases the intelligibility of the high-throughput data generated by ‘omics’ technologies such as transcriptomics, metabolomics and proteomics**.**

There are challenges in bioinformatics such as identification methods to best represent and organize data, how the information coming from different sources and disciplines will be analyzed and combined, and how scientists can access and search this information [Hemminger et al., 2005].

There are two major bioinformatics approaches with focus on molecular interactions that can be applied to transcriptomic datasets; Subnetwork discovery and Network Inference. In the subnetwork discovery approach, transcriptome data are mapped on interactome data, which includes information about proteins and their interactions with each other. Subnetworks are separately identifiable parts of a larger network that are supposed to have important nodes and edges for a certain condition [Beisser et al., 2010]. Subnetworks are computationally discovered from large networks by using the score of each gene in the network, usually derived from its p-value. These subnetworks can provide functionally related genes since they consist of genes whose change occurs simultaneously in the given condition. There are several computational methods to discover subnetworks from PPI. Batra et al. [Batra et al., 2016] reviewed 19 network enrichment methods and compared 7 of them over one selected dataset (Table 2.1).

Table 2.1: De novo pathway enrichment methods.

Tool	Method	Software
BioNet*	ASO	App
ClustEx	Clust.	App
cMonkey	Clust.	App
COSINE*	SP	App
GiGA*	SP	App
GXNA*	SP	App
HotNet	SP	App
jActiveModules	ASO	C-PL
KeyPathwayMiner*	MC	app, C-PL, WS
DEGAS*	MC	App
MEMCover	MC	App
NetWalker	ASO	App
NetworkTrail	ASO	WS

Table 2.1: De novo pathway enrichment methods (continue).

Tool	Method	Software
PinnacleZ*	ASO	app, C-PL
ReactomeViz-MCL	Clust.	C-PL
RegMOD	SP	App
ResponseNet	ASO	WS
SubExtract	ASO	App
TieDIE	SP	script (Python)

* ASO: Aggregate Score Optimization, SP: Score Propagation, MC: Module Cover, Clust: Clustering Based. The availability of the tool as a stand-alone application (app), as Cytoscape plugin (C-PL), or as web service (WS) [Batra et al., 2017].

The study showed that the performance of two methods were better than others, (Figure 2.3) [Batra et al., 2017]. KeyPathwayMiner [Alcaraz et al., 2014] and BioNet [Beisser et al., 2010]. Therefore, these two algorithms were chosen to be used in this study.

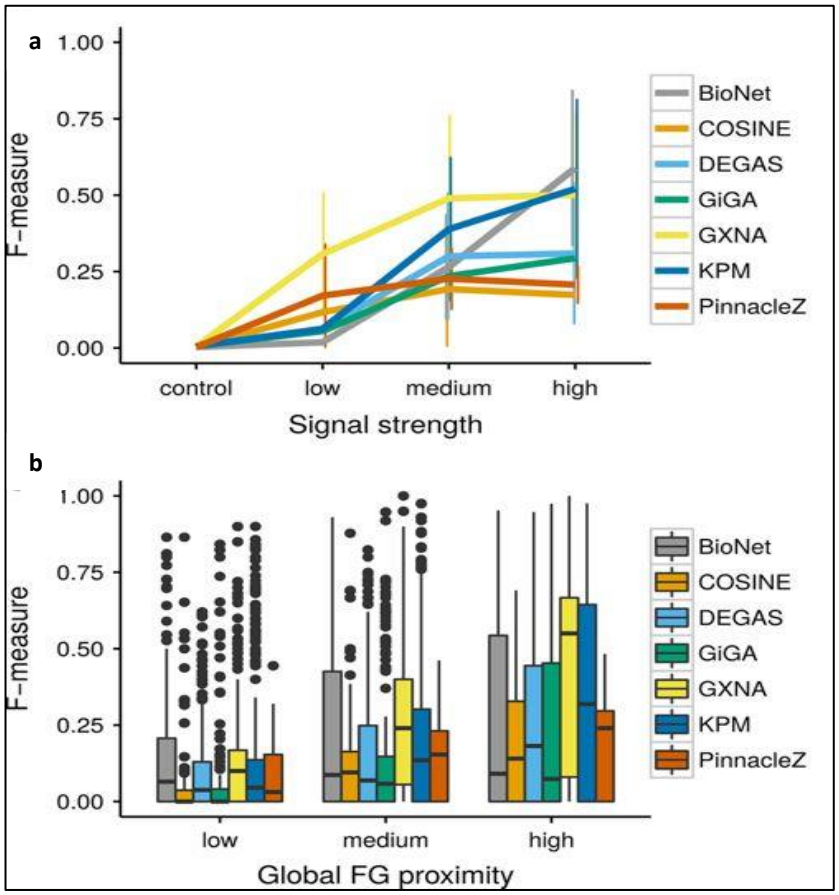

Figure 2.3: The performance of subnetwork analysis tool. The performance of tools based on the Signal Strenght (a) and Global FC Proximity (b) are shown.

2.5.1. KeyPathwayMiner

KeyPathwayMiner (KPM) [Alcaraz et al., 2014] is a Subnetwork Discovery tool that takes p-values and interactome data of an organism and gives subnetworks as output. This tool can also combine multiple omic studies and networks to create condition specific pathways and sub-networks. KPM needs two inputs: first one is p-values from the expression data and second one is organism-specific interactome data. KeyPathwayMiner accepts significance of change in a binarized format as input, where 0 shows no change, and 1 shows significant change for each gene. For the binarization, a threshold is required. This algorithm has a parameter called K, which shows how many non-significant genes are allowed in the subnetwork. It can take values from 0 to 10.

KPM algorithm can be used via three different platforms: Standalone, online usage in website (Web 1, 2018) and in Cytoscape as an add-on. If the PPI networks are huge, standalone version is more advantageous in terms of computational time. Cytoscape version can have problems in importing huge PPI data, or it may take long time. On the other hand, the Cytoscape version has an advantage of visualizing subnetworks and comparing them, and this version allows user to choose number of output subnetworks. Also, networks can be analysed and their size, number of nodes and edges, their most connected genes (hub), node metrics such as betweenness centrality or degree can be visualized by Cytoscape. Online usage is the slowest method and it does not work well for huge PPI data, but the nodes are linked to NCBI. If user clicks on any gene on the subnetwork found by KPM, the properties of the gene (node) such as node name, definition, organism type can be shown in the online version, which are retrieved from the NCBI.

There are studies that used KPM approach to analyze transcriptomic data. The molecular mechanisms of Huntigton's Disease were investigated by using KPM [Alcaraz et al., 2011], and the effect of increasing K on the subnetwork size was tested. The behaviour of huntingtin protein in the subnetwork was also analyzed to understand Huntigton's Disease mechanisms. KPM was used to find biomarker for the time series of tumor cells [Huang et al., 2017], and the relation between four tumor types were investigated through identified subnetworks. Chemotherapy response of breast cancer was also analyzed with KPM algorithm. That study aimed to find significant modules. These modules included genes that are significantly active and expressed in breast

cancer condition. The genes that were not in modules but significantly changed were identified as inactive genes without any effect on disease condition [Warsow et al., 2013].

2.5.2. BioNet

BioNet is another method to find subnetworks by using PPI and p-values of genes [Beisser et al., 2010]. BioNet also needs two datasets: interactome and transcriptome data. P-values are obtained from the transcriptome data, and each node (gene) is scored by using its p-value. The aim of this method is to find maximal scoring sub-graphs or, in other words, to find functional sub-networks from the large PPI data. The method is available through an R package with the same name. This package can be downloaded to R from Bioconductor [Web 2, 2018]. Visualization can also be done with R, but networks can be exported to other visualization tools like Cytoscape.

Generally, R functions require the input files in specific formats. Therefore, format conversion can be needed. For example, the interactome data was converted into GraphNel format from text format. This format is useful while working with graphs. Node and edge information can be stored in GraphNel format. This format also facilitates visualization in R.

BioNet does not changes p-values and uses them while scoring nodes with scoreNodes function. This function has one parameter called false discovery rate. This parameter is used as a threshold while scoring nodes. FDR means the expected proportion of false discoveries [Dittrich et al., 2008].

The method was used in lots of studies in literature. Ovarian cancer was analyzed by using BioNet. In this study, 74 ovarian cancer cases and 47 healthy controls were compared to find differentially expressed genes and these genes were mapped on interactome data to find significant modules by using BioNet [Yin et al., 2016]. Cheng et al used BioNet to predict recurrence risk of ovarian cancer in patients. In this study RNA sequencing data was analysed and related genes were aimed to be identified for the early detection of ovarian cancer [Cheng et al., 2018]. In another study, bladder cancer was investigated by silencing Tak1 gene, and gene expression data was used to create subnetworks with BioNet to see the effect of silencing [Chen et al., 2017].

2.6. Correlation Based Analysis of Transcriptome

Network inference is another bioinformatics analysis approach based on molecular interactions. Thanks to new omics technologies, thousands of genes and their behaviors in the cells can be observed. To analyse this high-throughput data, bioinformaticians need to combine some analysis techniques such as graph analysis, dynamic modeling, statistical inference etc. Pearson correlation is one of the most commonly used algorithms to find relationships of gene pairs. Gaussian Graphical models (GGM) is another algorithm that are used in network inference approaches. GGM allows the control of one variable to investigate the effects of other variables on each other. Thus, the net relationship between the two variables is revealed **[**Werhli et al., 2006], [Albert et al., 2007].

2.6.1. Co-expression analysis by Pearson correlation

Genes can be clustered based on the similarity in their expression profiles. Correlation methods provide information about groups of genes that react similarly to changing conditions and therefore are potantially co-regulated. However, the fact that the two nodes are in the same group does not mean that there is a causal relationship between them.

There are several methods to calculate correlations. Pearson correlation is one of the most commonly used computational methods. This method involves calculating the standard Pearson correlation coefficient (PCC) between expression values of gene pairs. PCC is a metric that scores the tendency of two genes to show similar expression profiles between samples. The PCC ranges from -1 to 1 because this relationship between genes can be in same or opposite direction. As a result, the absolute value of PCC is widely used as a measure of similarity ranging from 0 to 1. To make sense of these correlation values, a threshold must be selected. Based on that threshold gene pairs can be defined as co-regulated or not. In transcriptome data, the correlation between genes is calculated using gene expression values, and Pearson correlation creates a co-expression gene network [Krumsiek et al., 2011].

Pearson correlation is used to analyse biological data in literature. Llinás et al used Pearson correlation to understand the expression differences of lab strains and wild strains of *Saccharomyces cerevisiae.* At the end of the analysis, they discovered

that, three strains of *Saccharomyces cerevisiae* have different drug sensitivity [Llinas et al., 2006]. In another study, macrophage differentiation was analyzed with different types of stimuli. Diverse set of stimulants was applied to the human macrophages, and transcriptional regulation of macrophages was compared by using co-expression networks that are created by Pearson correlation. Central transcriptional regulators were identified [Xue et al., 2014].

2.6.2. Weighted Gene Correlation Network Analysis

One of the commonly used bioinformatics approaches for transcriptome data is correlation network analysis. Correlation networks are used for high-dimensional, large biological data to understand relationships of genes with each other. These networks can provide simple representation of nodes and their relationship. WGCNA [Langfelder et al., 2008] is one of the most famous methods to find networks from the correlation of genes. These methods provide simple representation of genes and their relationships, connections or systems-level functionality [Zhang et al., 2005]. Co-expression matrices are created, and each edge is weighted with their correlation value. Then, co-expression information is used to find modules that include highly correlated genes. WGCNA is an R package and freely available in Web 3.

There are several studies in literature that used WGCNA to find correlated networks. Horvath et al, [Horvath et al., 2012] used blood tissue and Human brain samples to find aging effects on DNA methylation. DNA methlation samples were assumed to be better than gene expression data, and this data was used to find highly correlated modules that are related to aging. In another study, comprehensive mRNA transcriptome data of multiple cancers, gastric and breast cancer were analyzed to detect varying stem cell features by using WGCNA [Kalamohan et al., 2019]. Another study used WGCNA to understand normal aging deficits and changing mechanisms in brain by comparing Alzheimer's disease gene expression data with normal aging data [Miller et al., 2008].

3. NETWORK BASED ANALYSIS OF MEMORY-ASSOCIATED TRANSCRIPTOME: DATASET I

3.1. PPI for *Rattus Norvegicus*

In this study, the aim was to create a rat-specific interactome that will be unique and includes all up-to-date interaction information in literature. Therefore, it was created by using five different interactome databases; BioGrid, Mint, Intact, Uniprot and iRefindex. From PSICQUIC, PPI information is downloaded in psimitab 2.5 format text file. The interactome was downloaded in December 2018. In these text files, among others, the information on interactions of proteins, the PUBMED ID's of articles that reported these interactions, interaction detection methods and organism names are reported. By using Miscrosoft Excel, organisms other than *Rattus norvegicus* and their interactions were removed from the downloaded list. Then, the gene symbol was chosen as the interactor name and if there is no gene symbol for that interactor, that interaction was removed. Genetic interactions and non-physical interactions identified by co-localization information were also removed from the data. Finally, all databases were aggregated, and self-loops and duplicated edges were removed by using Cytoscape [Shannon et al., 2003]. Final rat-specific unique interactome, called *Rattus norvegicus* interactome, had 3704 genes and 7304 interactions.

The goal in this part of the study is to map transcriptome data on interactome data. When the data was mapped on the interactome, the percentage of matching interactome was too low; 36% of the transcriptome data was mapped on the network and 55% of the genes in the interactome data had a mapping from the transcriptome data. Because there is a low percentage of mapping between transcriptome data and interactome data, many genes that may be relevant to memory and cognition were lost. As a solution to improve the interactome, the homolog genes between *Rattus norvegicus* and its close relative *Mus musculus* were searched by using Ensemble BioMart [Zerbino et al., 2018] (Figure 3.1). *Mus musculus* is phylogenetically the most related mammalian organism to *Rattus norvegicus*. They have a high number of common homolog genes. The following steps were followed here: (i) *Mus musculus* interactome was created by using the five databases mentioned above (December

2018), and the size of the interactome was identified as 10377 proteins and 36380 interactions. (ii) Homolog genes were found by Ensemble BioMart, which matches genes of different organisms based on sequence similarity. Ensemble Genes version 95 was chosen as the database from the website [Web 4, 2018], using the "Choose Database" link, and Rat genes (version Rnor_6.0) were chosen as the dataset. Then, from the attributes link on the left menu, "Homologues" option was selected because the goal is to get homolog genes between *Rattus norvegicus* and *Mus musculus*. From Gene section, "gene name" was selected, and from Orthologous section "mouse gene name" was selected from under "Mouse Orthologues" subsection. The result was obtained by clicking Results button on the upper left menu (Figure 3.1). BioMart gives 59.214 non-unique gene match. For this study, only the orthologous genes that have the same gene name in both organisms were selected to create interactome. The assumption here is that, given they are orthologs, if the genes have the same name in both organisms, they have a higher chance to carry the same function(s). There were 15.775 genes serving this purpose. The interactome data contains both gene names and protein names. In this thesis study, gene name is chosen for further analysis because transcriptome data was given in terms of gene names. The proteins coded by those genes have 30.607 interactions in *Mus musculus* interactome. Afterwards, the Rat interactome and this new homolog interactome were combined, and the number of common nodes in the constructed Rat interactome and *Mus musculus* interactome were found to be 3020. By using Cytoscape, the combined interactome was made unique, and a new bigger interactome was created. The final interactome, called RatMusInt, had 9500 genes (nodes) and 37043 interactions (edges). All interactome were created by using gene names as node names. Some characteristics of the interactomes mentioned here are given in Table 3.1. Visualization of this network was performed in Cytoscape and shown in Figure 3.2.

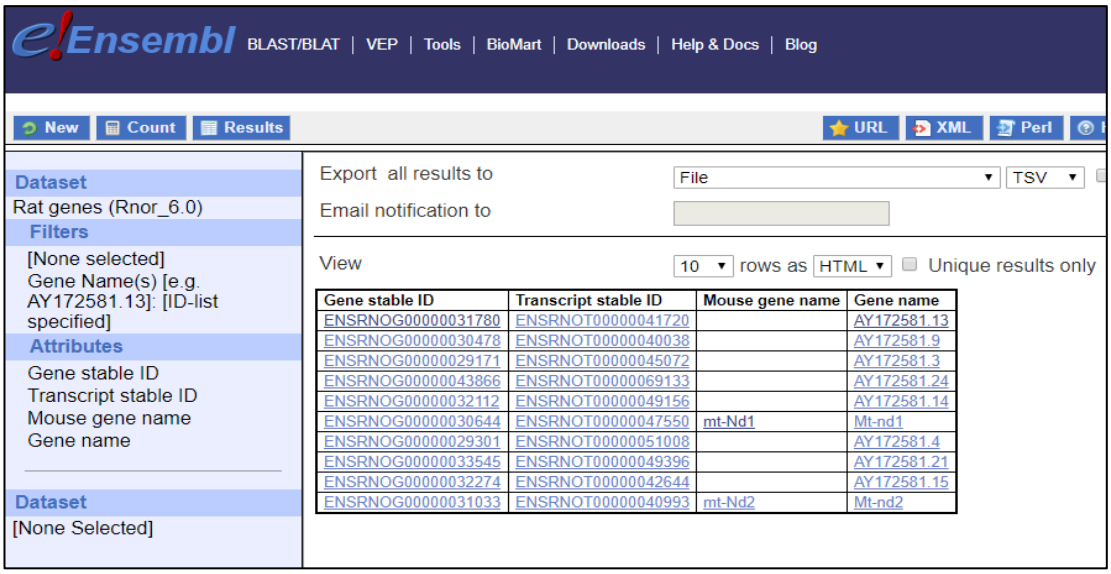

Figure 3.1: Ensemble- BioMart usage. Ensemble BioMart was used to find homolog genes between *Rattus norvegicus* and *Mus musculus.*

Table 3.1: Characteristics of interactomes. In this table, there are 3 interactomes with their number of nodes and edges. *Rattus norvegicus* interactome was created by using only *Rattus norvegicus* proteins, *Mus musculus* interactome was created by using only *Mus musculus* proteins and Ratmusint created by combining *Rattus norvegicus* interactome with the interactions of homolog *Mus musculus* proteins. Highly connected (hub) proteins and their number of interactions are also given in the table.

Rattus norvegicus interactome (3704 node -7304 edge)	*Mus musculus* interactome (10377 node-36830 edge)	Ratmusint interactome (9500 node-37043 edge)
Highest degree nodes Ubc: 825 Slc2A4: 609 Itm2B: 512	Highest degree nodes Fancd2: 1671 Eed: 1209 Ubc: 772	Highest degree nodes Fancd2: 1435 Ubc: 1398 Eed: 1054

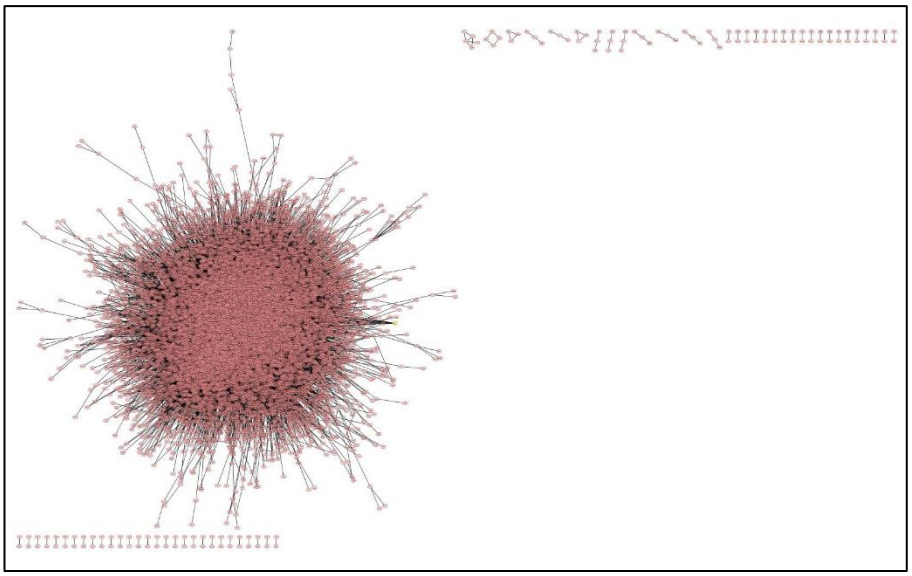

Figure 3.2: Whole PPI network for *Rattus norvegicus.*

3.2. Dataset Description

In literature, there are many different experimental designs that use different age groups and use memory and cognition tests to understand important mechanisms and pathways for learning and memory in aging brain. Blalock et al. (Blalock et al. 2003) created an experiment by using 4 (young) -14 (middle aged)-24 (aged) month-old rats to understand the effect of aging in memory and learning performance. Rats were trained for 7 days to see their physical performance in accomplishing Morris Spatial Water Maze test, in which rats are trained to escape from a tank. That tank is filled with water and there is a platform in it, on which rats can stay to escape from the water. Velocity and latency of rats in water can be measured and used to understand memory performance. Object Memory Task is another test that detects the performance change by aging but less dependent on physical strength. Here, some objects are placed in a box and rats can examine them, then a new object is added to the box. It is expected that rats will spend more time to examine the new object than already existing ones. Object Memory Task was applied to the rats to see the cognition difference in age groups. There were 10 rats for each group. Transcriptome analysis of the samples from the rats was performed by microarray technology (Blalock et al. 2003). Transcriptome data was extracted from the hippocampal CA1 region of the brain and this data can be obtained from the Gene Expression Omnibus (GEO). GEO is a database that stores high-throughput gene expression and other functional

genomics data sets. The data was downloaded from GEO database (Clough et al 2016) with the accession number GSE854. Data can be accessed with its title or accession number in GEO. Some statistical analyses (detecting differential expressed genes) can also be done online through GEO.

In their study, transcriptome data was normalized by using z-score method and ANOVA was used to find differentially expressed genes. These genes were functionally analysed with Gene Ontology (GO). Based on the functional analysis results, aging and cognition related genes and mechanisms were predicted.

3.3. Preprocessing of Dataset

3.3.1. Dataset Normalization

In GEO database, data are stored with and without normalization. In this study, the preference is to process all data using the same normalization procedure. Therefore, raw data was used as input to RMA normalization algorithm to normalize expression values of each gene. Normalization is an important step in microarray-based transcriptome data to make relative intensities on the arrays comparable. Gene expression values are stored in CELL file as a raw data. Raw data in this format was downloaded from GEO, and the normalization processes was done with RMA package in R. RMA package has the function *rma* that converts an AffyBatch object into an ExpressionSet object using the robust multi-array average (RMA) expression measure.

3.3.2. PCA/Sammon mapping to Detect Outliers

GSE854 data consisted of 30 samples in total. To see the distribution of these samples with respect to each other in terms of their gene expression values and to find if there were any outliers, we performed PCA analysis and then Sammon mapping in MATLAB. PCA is an unsupervised dimension reduction method. The aim is to project transcriptome data in 2 or 3-dimensional space. Thanks to PCA, samples are compared in 2 or 3 dimensions rather than approximately 20.000 dimensions (each gene with measurement values across samples constitutes a different dimension). In PCA

analysis, grouping is done by reducing dimension. The behavior of each sample can be visualized in this way, and outliers can be detected if any.

In the first step of PCA, data is normalized by auto scaling. The mean of each gene is calculated and extracted from each gene expression (M). The standard variation (S) is calculated for each gene. Autoscaled data is created by M divided by S (data=M/S). In this way, all the genes in the data will have a mean of zero and a standard deviation of one across samples in the autoscaled data. This process is done to make all variables on the same scale. The Covariance matrix (C) is then calculated by applying *cov* function in MATLAB to the data matrix. Eigenvalues and Eigenvectors of Covariance matrix are later calculated by *eig* function in MATLAB. The biggest eigenvectors are chosen to represent the direction with highest variances. Loading and Scores matrices are determined, and scores plot is created. Also, there is the *pca* function in MATLAB that can do all these processes automatically.

With PCA analysis, the grouping of samples can be observed while samples from many different groups are considered, but in Sammon mapping only two groups of samples can be compared. Therefore, if only the relationships between two groups are examined, Sammon mapping can be used. Sammon mapping analyses were also done in MATLAB. First, the groups are introduced to MATLAB separately. With *pdist* function, Euclidean distance between pairs of observations is calculated. *mdscale* function performs nonmetric multidimensional scaling on the *n*-by-*n* distance matrix, and returns Y, a configuration of *n* points (rows) in p dimensions (columns). Then *mdscale* results can be plotted for two conditions and their relationship can be observed.

PCA analysis was performed for all samples together, and Sammon mapping analysis was performed seperately for binary comparison of each age group. Based on the PCA results for all samples (Figure 3.3), Middle aged 8 and Young Sample 4 and Middle-aged Sample 8 are outliers, they are both far away from the other samples of the respective age groups. Aged samples are grouped together, there were no outlier for this age group. Therefore, Sammon mapping was performed for Middle aged and Young samples to understand their behaviour together. In sammon mapping only Middle aged 8 was located far from the other samples of the group. Middle aged sample 8 was an outlier in both analyses, and it was removed from the dataset (Figure 3.3 and 3.4). Further analyses were performed with 29 samples.

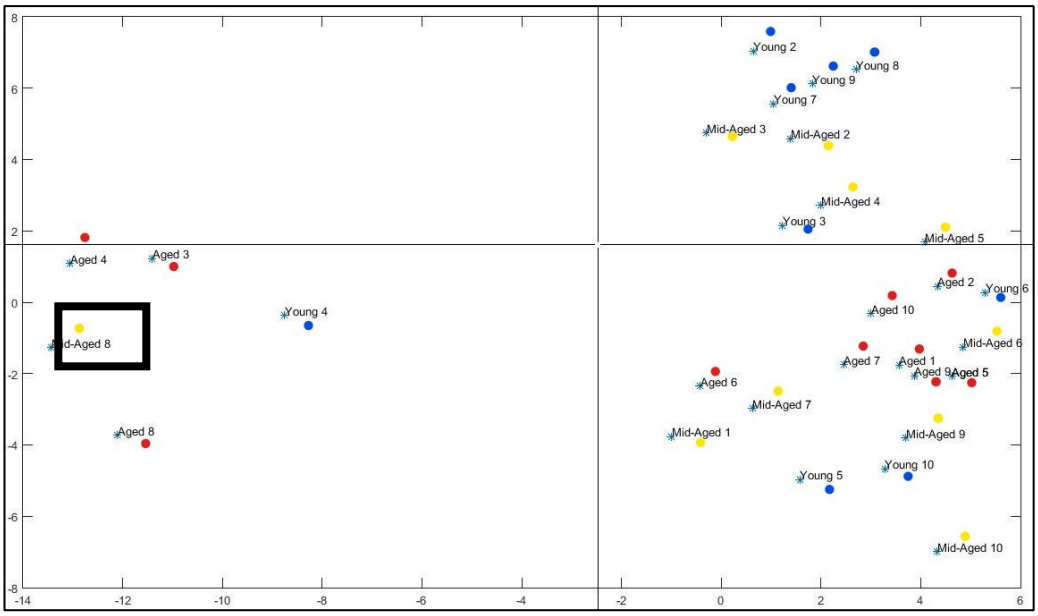

Figure 3.3: GSE854 PCA result.

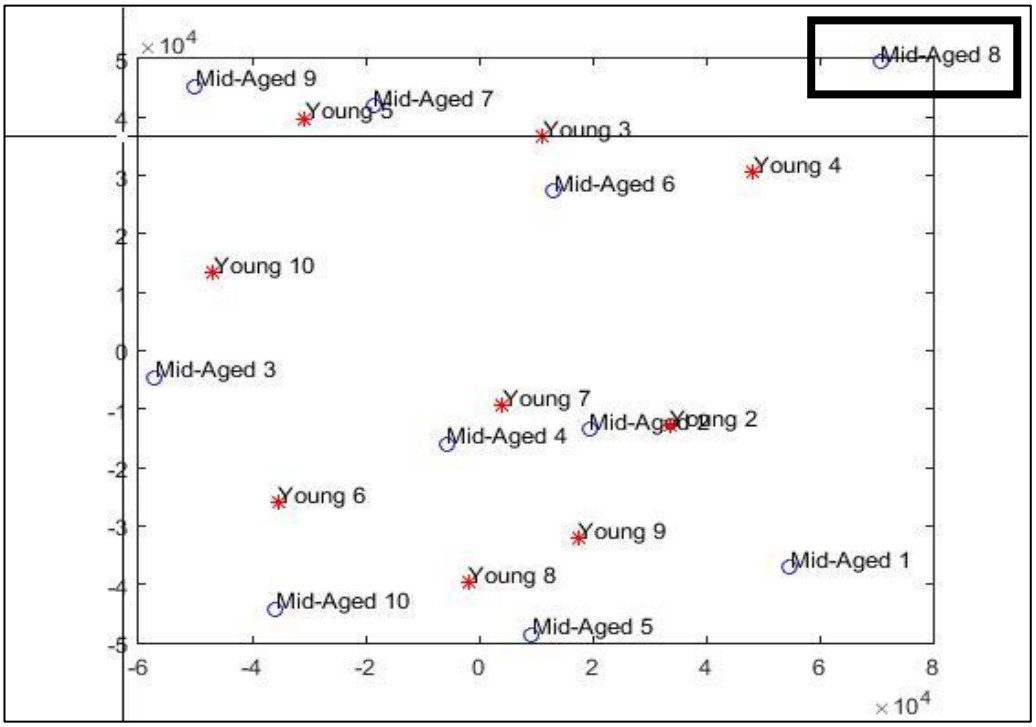

Figure 3.4: GSE854 Sammon mapping result.

3.3.3. Calculation of P-values

The third process after normalization and PCA-Sammon mapping is the calculation of p-value. The p-value of each gene is calculated and used as score for each gene. (Beisser et al. 2010). Gene names/IDs and corresponding p-values are

needed for further analysis. RMA-normalized data was used to calculate p-values. Limma package (R) was used for this purpose. Limma is an R package for the analysis of gene expression microarray data by using linear models. With Limma, many RNA targets can be compared simultaneously. Limma uses Empirical Bayesian method and it provides stable results even for small number of arrays. Contrary to standard t-test, limma considers all genes while calculating p-values. There were three age groups in the dataset. We created binary combinations between them to calculate differential p-values: (i) Young versus middle aged, (ii) Middle aged versus aged, (iii) Young versus aged.

The variance of expression measurements on many platforms (arrays, etc.) depends on the expression level. By log-transforming, dependence is reduced, and data gets closer to normal distribution. With log2 transformation count data can be scaled. The data should be transformed using log2 for Limma-based p-value calculation. Based on Limma results, p-values of 8800 genes were obtained. P-values can be obtained from the result of Limma with or without correction. For this dataset, p-values, without correction, were used with 0.01 and 0.05 cut-off. We used gene symbols to represent a gene, and some genes did not have a symbol in the dataset, they had only probe IDs. We removed genes that did not have gene symbols (951 genes). In the transcriptome data some genes have more than one symbols (529 genes) that are separated from each other by "//". They are represented by one probe and one expression value in the data. For this study, all gene symbols were decided to be used. For example, if probe 1 has four genes in the following format AB//AB1//ABC//ABC1, and the expression value of this probe is 8.5, this value is assigned to each symbol seperately and dedicate independent lines in the transcriptome data for each of such symbols. There were more than one probes for some genes, and the most significantly changed probe is used to represent such genes based on their p-value. Thanks to this pre-processing step, all genes and their p-values were made unique. 5749 gene symbols and their p-values were used for further analysis.

3.4. KeyPathwayMiner (KPM) Analysis for Dataset I

KeyPathwayMiner (KPM) is a condition specific pathway analysis tool that takes p-values and interactome data of organism and gives subnetworks as output. A

p-value threshold of 0.01 was used mostly in this study. If the number of significantly changed genes was too low at 0.01 threshold, 0.05 was used as threshold. If p-value of a gene was above threshold, it was as assigned a value of 0, if not; it was assigned 1. In this thesis study, K=0, K=1 and K=2 were used. If the size of the subnetwork is too low –the number of nodes is lower than 20 that is not preferred, and a higher value of K can be used.

KPM needs two inputs: interactome and transcriptome. We introduced our "ratmusint" interactome data to Cytoscape. In this study Cytoscape was used for KPM analysis. Preprocessed transcriptome data was also given as an input to the KPM. Then, subnetworks were identified for different K values. 61% of the transcriptome is mapped on interactome, and 37% of the interactome is mapped on transcriptome data. This means nearly 3506 genes were mapped on the interactome. Each subnetwork was saved, and their topological analyses were performed. (Table 3.2).

KPM results are shown in Table 3.1. Subnetworks were created for different K values. K was changed from 0 to 2. K value of 0 means a subnetwork with only significant genes. K of 2, on the other hand, allows the inclusion of maximum 2 non-significant genes in the subnetwork. As shown in Table 2, we can see different hub genes in the identified subnetworks when different K values are used. For example, in Young- Middle aged analysis at K=0, Slc2a4 is hub gene, namely it has the highest number of edges in the subnetwork. When we changed K to 1, the size of the network increases, and hub gene was changed to Ubc. In general, if nonsignificant genes are allowed in network, KPM includes the most connected (highest degree) non-significant gene in the subnetwork for K=1. That gene then becomes the new hub gene in the subnetwork. We checked K=2 case also. When two non-significant genes were allowed, KPM again included the two highest-degree non-significant genes into the subnetwork and, therefore, the size of the network increased.

Dataset I (GSE 854) contains three age groups and pairwise comparison of each age group was analyzed by KPM. At the first column of Table 3.2, we can see the number of significant genes with respect to two different thresholds. Results corresponding to the cut-off value that are shown in bold in the table were used for KPM analysis. The cut-off selection is determined based on the number of diffentially expressed genes for that cut-off. Also, we can see the compared conditions and number of samples for each condition in the table. For this dataset, functional analysis of the genes in the subnetwork was performed for K=0 to detect which common properties

these significantly changed genes have. At K=0, the size of the subnetworks is acceptable for each comparison group.

The number of nodes that were identified with KPM analysis should be between 50-150 for functional analysis. A smaller subnetwork may not reflect commonly functioning gene groups, and a larger subnetwork may lose specificity. Therefore, K = 0 was chosen. Visualization of subnetworks was performed in Cytoscape. Young-Aged subnetwork is shown in Figure 3.4.

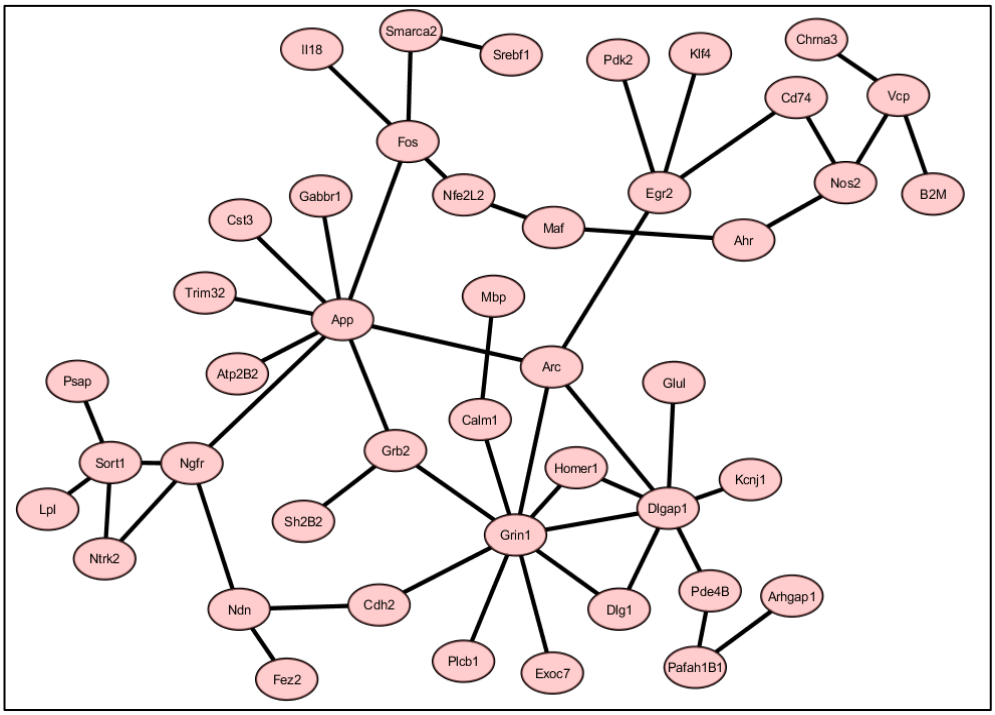

Figure 3.5: Young-aged comparison subnetwork obtained by KPM for Dataset I for K=0

Table 3.2: KPM results for GSE 854 datasets.

Age group: Explanations of data:	K=0	K=1	K=2
	Size of networks and highest and second highest degree nodes		
Young (9) - Middle aged (9) **p<0.05=855** p<0.01=276	231 node-365 edge Slc2a4: 82 Grin2b: 35	313 node-533 edge Ubc:147 Slc2a4: 83	350 node-651 edge Ubc: 148 Fancd2: 118
Middle aged (9) - Aged (10) **p<0.05=881** p<0.01=216	196 node-264 edge Mapk3: 29 Grin1: 21	259 node-369 edge Mapk3: 29 Grin1: 21	301 node-481 edge Ubc: 111 Fancd2: 98
Young (9) - Aged (10) p<0.05=1085 **p<0.01=407**	44 node-50 edge Grin1: 9 App: 8	83 node-94 edge Fancd2: 42 Grin1: 9	106 node-156 edge Ubc: 62 Slc2a4: 43

3.4.1. Functional analysis of KPM subnetworks for Dataset I

Enrichment tests are applied to the genes from the subnetworks to understand the function of these subnetworks. g:Profiler [Reimand et al., 2016] was used for functional annotation. This tool takes a gene list and finds statistically significant Gene Ontology terms, pathways and other gene function related terms. GO can be defined as a controlled vocabulary where the roles of genes and proteins in the cell are defined. It is divided into three categories; Biological process (in which biological processes is that gene or protein involved), Molecular function (what is the function of that protein or gene), and Cellular component (to which cellular part does that gene or protein belong) [The Gene Ontology Consortium, 2019]. From the options section, p-value threshold for significance can be adjusted. For this analysis p-value threshold was chosen as 0.05. And, significance threshold link allows us to choose the type of correction. There are three correction types; Bonferroni correction, Benjamini-Hochberg correction and g:Profiler correction. For the analyses in this thesis, the default correction type was used (g:Profiler correction). The output type can be chosen as graphical, textual format or EXCEL spreadsheet. We can also obtain enriched transcription factors that are related to our genes and enriched protein complexes from

CORUM database [Giurgiu et al., 2019], which stores manually curated protein complexes that were experimentally characterized (Figure 3.6).

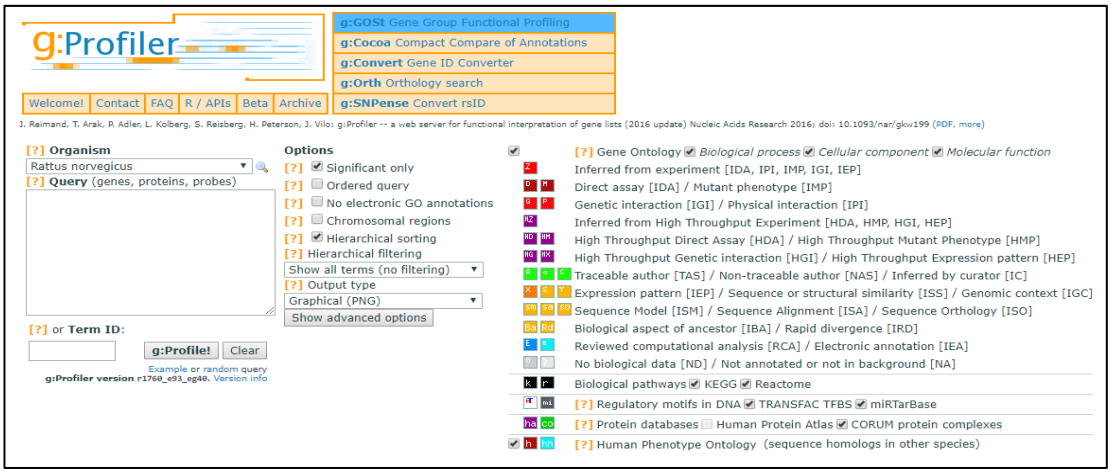

Figure 3.6: g: Profiler usage

Functional analysis was performed by using the genes of the identified subnetworks. To see if the identified subnetworks are relevant, we have given importance to the existence of certain terms in the functional analysis results, and we have given precedence to those terms. These terms were Learning, Memory, Plasticity and Nervous System, Neurogenesis. Blalock et al., aimed to find relationships between age and cognition deficits and the effect of aging process on learning and memory mechanisms. In this thesis, the aim is to find such relationships by using transcriptome data and interactome data. Plasticity, learning and memory are the terms that are expected to be identified in the functional analysis results because the focus in this study is to understand underlying mechanisms of cognition deficits by aging. Neurotrophin signaling pathway, which is important for neuron development, survival and function, is also detected in the functional analysis results. Brain-derived neurotrophic factor has a function in neuronal growth, synaptic and structural plasticity and morphology. This molecule arranges complexity of dendrides, long-term potentiation, axon branching, processes that affect synaptic efficacy and improves memoy and learning performance by creating synaptic connections between neurons [Mitre et al., 2017].

In this dataset, all comparison groups (Young vs. Middle aged, Middle aged vs. Aged and Young vs. Aged) had at least one of these terms. These subnetworks proved that the results of KPM are reliable, and the algorithm detects the modules that

includes the processes affected from the aging, by using whole transcriptome data. Tables 3.3-3.5 shows the functional annotation results for Dataset I at K=0. Only the memory, learning, cognition and nervous sytem related terms are given in the functional analysis results. All results are in the Suplementary Tables.

Table 3.3: g:Profiler analysis of KPM-based Young-Middle aged comparison. P-value refers to the significance of terms, term type refers to abbreviation of terms; BP: Biological Processes, CC: Cellular Component, keg: from Kegg databases and rea: from Reactome database, hp: information from Human Protein Atlas.

term type	term name	p-value
BP	neuron projection organization	0.0184
BP	response to cytokine	0.0023
BP	response to oxidative stress	1.87E-07
BP	response to reactive oxygen species	2.35E-05
BP	cell-cell signaling	5.34E-08
BP	cellular response to cytokine stimulus	0.00897
BP	nervous system development	0.00486
BP	Neurogenesis	0.011
BP	Gliogenesis	0.0178
BP	neuron projection morphogenesis	0.0343
BP	regulation of neurotransmitter levels	0.00227
BP	regulation of synaptic plasticity	6.71E-08
BP	regulation of neuronal synaptic plasticity	1.39E-08
BP	regulation of long-term neuronal synaptic plasticity	2.79E-05
BP	positive regulation of neuron death	0.0164
BP	Cognition	0.0171
BP	learning or memory	0.00582
BP	Learning	0.0489
CC	neuron to neuron synapse	7.73E-14
CC	neuron part	3.06E-21
CC	neuron projection	1.26E-14
CC	neuron spine	1.35E-05
CC	neuronal cell body	6.79E-06
Hp	Neoplasm of the nervous system	0.00446
Hp	Neoplasm of the peripheral nervous system	0.00076
Keg	Neurotrophin signaling pathway	0.000747
Rea	Neuronal System	1.37E-06
Rea	Neurotransmitter receptors and postsynaptic signal transmission	6.63E-05
Rea	Glutamate binding. activation of AMPA receptors and synaptic plasticity	5.88E-07

Table 3.4: g:Profiler results for KPM-based Middle aged-Aged comparison.

term type	term name	p-value
BP	Aging	1.79E-10
BP	circadian rhythm	0.000444
BP	inflammatory response	0.000295
BP	cytokine production	3.66E-15
BP	nervous system development	4.26E-15
BP	central nervous system development	1.9E-09
BP	neuron death	1.1E-13
BP	neuron apoptotic process	1.12E-10
BP	Neurogenesis	1.75E-14
BP	Gliogenesis	0.000000192
BP	generation of neurons	1.95E-14
BP	neuron differentiation	1.85E-13
BP	neuron development	1.06E-11
BP	cell-cell signaling	0.000000641
BP	regulation of neuron death	1.08E-12
BP	cytokine-mediated signaling pathway	0.0000136
BP	regulation of nervous system development	1.54E-08
BP	regulation of neurogenesis	1.03E-09
BP	regulation of neuron differentiation	5.7E-09
BP	regulation of gliogenesis	0.0319
BP	regulation of synaptic plasticity	0.00855
BP	regulation of long-term neuronal synaptic plasticity	0.000402
BP	positive regulation of neurogenesis	0.0000939
BP	positive regulation of neuron differentiation	0.00277
BP	Cognition	2.13E-09
BP	learning or memory	3.36E-10
CC	neuron part	2.56E-09
CC	neuron projection	6.28E-10
CC	myelin sheath	0.0000277
MF	cytokine receptor binding	0.00573
keg	Neurotrophin signaling pathway	7.24E-12
rea	Cytokine Signaling in Immune system	0.00045

Table 3.5: g:Profiler results for KPM-based Young-Aged comparison.

term type	term name	p-value
BP	cytokine production	8.87E-08
BP	response to cytokine	6.29E-05
BP	Signaling	5.36E-07
BP	Aging	0.00949
BP	nervous system development	1.06E-06
BP	cell-cell signaling	1.19E-06
BP	Neurogenesis	0.00117
BP	generation of neurons	0.000462
BP	Gliogenesis	0.0356
BP	neuron differentiation	0.000918
BP	neuron development	0.000636
BP	cellular response to cytokine stimulus	0.00162
BP	cellular response to oxidative stress	0.0225
BP	regulation of cytokine production	8.52E-05
BP	positive regulation of cytokine production	0.000604
BP	regulation of neuronal synaptic plasticity	0.00753
BP	regulation of nervous system development	0.000126
BP	regulation of neurogenesis	0.00043
BP	regulation of neuron differentiation	0.00059
BP	Cognition	0.000129
BP	learning or memory	6.04E-05
BP	Learning	0.0195
BP	neuron projection morphogenesis	0.0189
BP	immune system process	0.0177
CC	neuron to neuron synapse	5.19E-07
CC	neuronal cell body	7.86E-05
CC	neuron projection	3.67E-06
CC	intracellular vesicle	0.0281
CC	cytoplasmic vesicle	0.0274
MF	calmodulin binding	0.000316
keg	Neurotrophin signaling pathway	0.0172
rea	Neuronal System	0.000568
rea	Neurotransmitter receptors and postsynaptic signal transmission	0.0402

3.5. BioNet Analysis for Dataset I

BioNet also needs two datasets: interactome and transcriptome data. Our interactome data was in binary text format, it means we have two columns, and edges are defined as “From” and “To” (from one gene to another gene). BioNet does not accept transcriptome data in text format, transcriptome format was modified. With the ftM2graphNEL function in BioNet package, the interactome data -includes source node and target node information- is taken from its binary matrix format and converted into GraphNel format.

The second data -transcriptome data- was also introduced to the R. This data includes gene names and their p-values. We did not make any change in p-values unlike KPM, where the data had to be binarized. With aggrPvals function of BioNet package, if there are p-values from heterogeneous experiments, they can be aggregated to get a single p-value for each gene. We did not change this parameter and fixed it to 1 for further analysis because there is a single p-value for each gene in this case. Then, gene names were mapped on the interactome data, and a variable called MappedNetwork was created.

FDR is one of the most important parameters in BioNet. This parameter is changed from 0.01 (most significant) to 0.2 (least significant) to find best possible subnetwork. FDR of 0.01 means that out of 100 significantly changed genes, the chance to have a false positive is 1. FDR of 0.2 means 20 out of 100 significant values can be false positives. While FDR increases, the significance of the result decreases. Therefore, an FDR value of higher than 0.2 was not used in this study. runFastHeinz function of BioNet creates subnetworks in GraphNel format, which can be saved and used for further functional analyses. Subnetworks and their properties are shown in Table 3.3 for Dataset I. Definitions of groups were given in the first column of the table. MappedNetwork shows the information of genes in transcriptome data that were mapped on interactome data. Nearly 53% of the transcriptome was mapped on interactome. Mapped Network has 3453 nodes and 11161 edge. Ubc is one of the mostl connected (hub) genes with 700 edges, and Eed follows it with 433 edges. Sample size for each group is also given in the table. The number of significantly changed genes for each comparison is also available, together with the size of each subnetwork and their hub genes. Unlike KPM, the hub genes do not change when the size of the

network changes. Visualization of subnetworks was performed in Cytoscape and Young-Aged comparison subnetwork is shown in Figure 3.5.

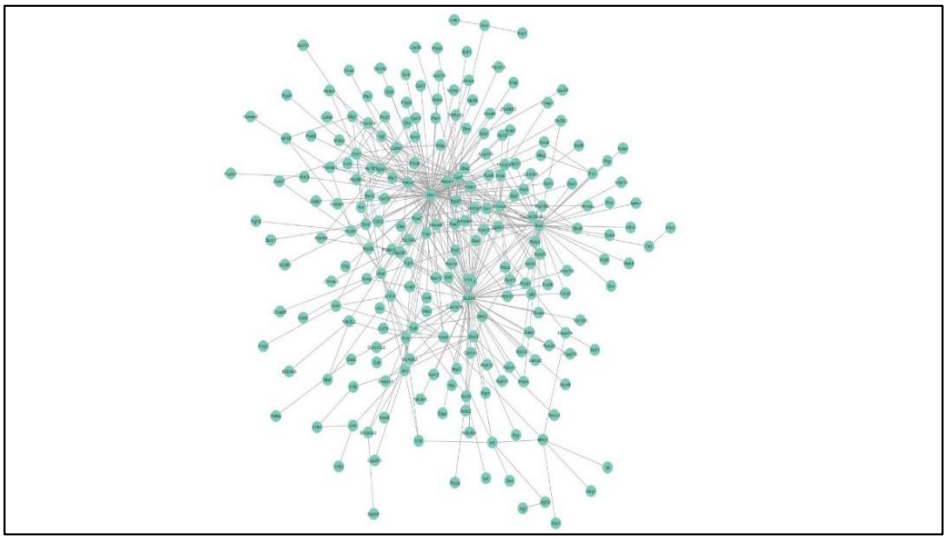

Figure 3.7: Young-Aged comparison obtained by BioNet for Dataset I at FDR of 0.1

Table 3.6 BioNet results for GSE854 dataset.

Bionet	FDR =0.05	FDR =0.1	FDR =0.2
	Size of networks and highest degree of nodes		
Young (9) – Middle aged (9) p<0.05=855 p<0.01=276	38 node-57 edge Ubc: 22 Itm2b: 19	113 node-178 edge Ubc: 58 Itm2b: 41	*
Middle aged (9) – Aged (10) p<0.05=881 p<0.01=216	No network was found	58 node-70 edge Ubc: 26 Slc2a4: 22	379 node-685 edge Ubc: 125 Slc2a4: 82
Young (9) – Aged (10) p<0.05=1085 p<0.01=407	68 node-101 edge Ubc:42 Slc2a4:26	219 node-419 edge Ubc:91 Slc2a4:64	*
*sub-networks were not analysed at this FDR. The significance of modules is low at FDR=0.2, so if the number of nodes and edges are enough- higher than 100 nodes- at FDR=0.1, subnetworks were not determined at FDR=0.2.			

3.5.1. Functional analysis of BioNet subnetworks for Dataset I

Functional analyses were done by using g:Profiler, as mentioned in Chapter 3.4.1. Young-Middle aged and Aged- Young comparison were done at FDR=0.1, Middle aged-Aged comparison was done at FDR=0.2. Based on the functional analysis, BioNet and KPM pointed to nearly same biological pathways, but KPM modules have more relavant terms than BioNet modules. All three analyses of KPM include Memory, Learning and Plasticity terms, which are important for aging mechanisms. On the other hand, two of BioNet modules include these terms. Myelination, which is thought to be an important mechanism in brain aging [Rowe et al., 2007], is detected only in BioNet modules in Middle aged-Aged comparison. Young-Middle aged comparison includes less terms for both algorithms in functional analysis results than the other two comparisons.

For these subnetworks, functional analysis were performed with g:Profiler correction method and p-value threshold was chosen as 0.05. Only the learning, memory, nervous system related terms are given in the results. Neurexins and neuroglins are important signaling pathways responsible for synapse development, these terms are also identified in the results. The trans-synaptic interaction between these two molecules is required for stability, plasticity, function, structure and creation of synapse which is important for memory and learning mechanisms [Brito-Moreira et al., 2017].

Table 3.7: g: Profiler results for BioNet-based Young-Middle-aged comparison.

term type	term name	p-value
BP	cell-cell signaling	0.00572
BP	response to oxidative stress	0.0169
BP	regulation of synaptic plasticity	0.049
BP	regulation of neuronal synaptic plasticity	0.000526
CC	neuron part	9.81E-08
CC	neuron projection	2.92E-06
CC	neuron to neuron synapse	0.00227
Rea	Neuronal System	6.62E-05
Rea	Neurotransmitter receptors and postsynaptic signal transmission	0.00409
Rea	Neurexins and neuroligins	0.0291

Table 3.8: g:Profiler results for BioNet-based Middle-aged-Aged comparison.

term type	term name	p-value
BP	circadian rhythm	0.0141
BP	cell-cell signaling	3.49E-08
BP	cell aging	0.0105
BP	neuron death	1.77E-13
BP	cytokine production	8.94E-11
BP	Myelination	0.00895
BP	Neurogenesis	3.2E-14
BP	Gliogenesis	3.66E-08
BP	neuron differentiation	3.76E-11
BP	central nervous system development	2.39E-13
BP	cellular response to cytokine stimulus	1.71E-10
BP	inflammatory response	0.0214
BP	response to oxidative stress	2.8E-16
BP	neuron projection regeneration	0.00088
BP	regulation of cytokine production	2.48E-10
BP	regulation of long-term neuronal synaptic plasticity	0.000849
BP	Cognition	3.36E-08
BP	learning or memory	2.56E-08
BP	regulation of cytokine production involved in immune response	0.0181
CC	neuron to neuron synapse	0.00814
CC	neuron projection	7.09E-12
keg	cAMP signaling pathway	0.00000435
keg	Neurotrophin signaling pathway	9.43E-09
rea	Immune System	0.00441
Rea	Cytokine Signaling in Immune system	0.0193
Rea	Innate Immune System	0.000516

Table 3.9: g:Profiler results for BioNet-based Aged-Young comparison.

term type	term name	p-value
BP	response to cytokine	3.23E-11
BP	regulation of neurotransmitter levels	0.00598
BP	cellular response to cytokine stimulus	3.52E-09
BP	cytokine production	1.17E-10
BP	Neurogenesis	0.0000961
BP	Gliogenesis	0.0000357
BP	generation of neurons	0.000414
BP	neuron differentiation	0.000917
BP	central nervous system development	8.71E-08
BP	Aging	0.00000287
BP	Behavior	0.0000645
BP	cell-cell signaling	6.67E-09
BP	synaptic signaling	0.00000229
BP	regulation of cytokine production	1.12E-08
BP	regulation of nervous system development	0.0000426
BP	regulation of neurogenesis	0.0000202
BP	regulation of neuron differentiation	0.000478
BP	circadian rhythm	0.00679
BP	positive regulation of cytokine production	0.0000152
BP	Cognition	0.000226
BP	learning or memory	0.0000617
BP	regulation of synaptic plasticity	0.0155
BP	cytokine-mediated signaling pathway	0.00135
BP	neurotrophin signaling pathway	0.0000502
BP	positive regulation of cytokine production involved in immune response	0.00257
CC	neuron to neuron synapse	0.000333
CC	neuron part	0.00000707
CC	neuronal cell body	0.00514
CC	neuron projection	0.00000877
MF	neurotrophin receptor activity	0.0105
keg	Neurotrophin signaling pathway	0.0237

3.6. Correlation Analysis for Dataset I

Dataset I (GSE854) [Blalock et al. 2003] was used to create a co-expression network. There were 8800 gene probes in the normalization results, but some probes did not have a gene name, so they were removed (1485). There were 7315 probes left, but there were some probes with the same gene name. To have a unique gene set with only one gene expression data for one gene, duplicate genes were removed by keeping the one with maximum gene expression values. There were 4873 genes left after this step. Further analysis was done by using this data. The number of genes that do not have name isdifferent from Section 3.3.3. When Limma analysis is performed, the names of the genes are searched on NCBI database and if there is a name available for that probe ID, it is added to the data. In correlation analysis, Limma was not used. Therefore, the number of genes that without an associated name was higher than in Section 3.3.2, and the number of genes used for correlation analysis was lower.

3.6.1. Co-expression analysis in Aged and Young group

Firstly, co-expression matrix for each age group was created. Young and Aged groups were imported into MATLAB separately and with "corr" function co-expression matrices were created. This function requires one input, that is normalized Dataset I (GSE854) data, and gives two outputs, co-expression matrix and p-value matrix. P-value matrix stores p-values of each correlation values. Co-expression matrix was used for further analysis. This matrix includes values from -1 to 1, corresponding to Pearson Correlation values. If the absolute correlation value between two genes is close to 1, this means that there is a strong correlation between these genes. Co-expression matrix consists of 4873 rows and 4873 columns for this data.

Correlations values can be either positive or negative. Both are important for this study. Threshold is needed to filter highly correlating gene pairs, and the threshold was chosen as 0.95. If the correlation value for a gene pair is higher than 0.95, that means they are significantly highly correlated for that condition.

Dataset I (GSE854) consists of three age groups, and each group has 10 samples. PCA result (Chapter 3.1.1) shows that sample Middle aged 8 was outlier and removed from the data. (Figure 3.2, 3.3). Table 3.4 reports results of correlation-based

analysis for aged and young datasets to find genes with distinct behavior in two conditions.

From the aged data, co-expression matrix was created for 0.95 threshold. This co-expression network was then imported to Cytoscape, and topological features were analyzed. These networks may consist of unconnected nodes and small networks. They can be removed from the main network to focus on a fully connected network. Extracted network for aged data has 2433 nodes and 23381 edges. Co-expression matrix was also created for Young data, and the extracted connected network has 1969 nodes and 3685 edges. The characteristics of these networks is shown in Table 3.10.

Table 3.10: Co-expression Networks for GSE854 dataset. These co-expression matrices were created for Aged and Young groups separately. Extracted networks represent the most connected network that was extracted from the initially constructed network.

Extracted aged	Extracted young
Size of networks and highest-degree nodes	
2433 nodes 23381 edges Cct3:176 Ddx3x:169	1969 nodes 3685 edges Ralgds:21 B4galnt1:18

3.6.2. Decreasing Co-expression Pattern Between Age Groups

In this part of study, we want to create a network pattern between age groups, to identify gene pairs that shows consistently decreasing correlation from aged to middle-aged to young and young to middle-aged to aged. Firstly, the correlation values between highly correlated 3685 gene pairs, that were calculated in Chapter 3.6.1 for young group, was also calculated in the other two age groups, middle aged and aged. The correlation value can be either positive or negative, so we separated this analysis into two groups; negative high correlation and positive high correlation. The correlation difference between each successive group was defined to be minimum 0.15 to have a pattern. This means that if the correlation value of one gene pair in Young age group decreases more than 0.15 in Middle aged, this correlation shows significant decrease, and obeys the rule (Table 3.11).

In negative correlation group, there were 306 gene pairs that showed consistent decrease in absolute correlation between the three age groups, from young to aged. Among them, 154 gene pairs obeyed the threshold rule. In positive correlation group, there were 635 interactions that were steadily descending. There were 232 interactions in this gene list that obeyed the rule. The two groups were combined after threshold filtering, and the combined network had 386 edges and 572 nodes. The network was analyzed in Cytoscape and mostly connected parts of the network were extracted. There were two unconnected subnetworks, which had 48 nodes-48 edges and 15 nodes and 14 edges. These two subnetworks were not big enough for further analysis, so we decided to change the threshold rule. A decrease of at least 0.1 in correlation was applied to the age groups for decreasing pattern. With the new rule, total positive and negative steadily descending genes and their interaction creates a network that has 744 nodes and 547 edges. The mostly connected sub-group of this network has 103 nodes and 108 edges.

Table 3.11: Co-Expression Network analysis for GSE854 dataset. Decreasing correlation pattern observed in Young to Aged and Aged to Young groups.

Network / Group	From Young to Aged	From Aged to Young
	Size of networks and highest degree of nodes	
Total network	744 node - 547 edge Gabrb1: 8 Prkg2: 6	1690 node - 4315 edge Ddx3x: 42 Ssr3: 42
Extracted Network	103 node 108 edge Gabrb1: 8 Prkg2: 6	1425 node 4154 edge Ddx3x: 42 Ssr3: 42

In negative correlation group, there were 3294 gene pairs that showed consistent decrease in absolute correlation between the three age groups, from aged to young. There were 1751 interactions in this gene list that obeyed the rule (correlation difference of 0.15 between age group). In positive correlation group, there were 6053 interactions with high correlation that were steadily descending. There were 2566 interactions in this gene list that obeyed the rule. The two groups were combined, and the total network had 1690 nodes and 4315 edges. The network was analyzed in

Cytoscape and mostly connected part of the network was extracted. This subnetwork has 1425 nodes and 4154 edges. Visualization of these networks was performed in Cytoscape. Decreasing correlation pattern for Aged to Young group extracted module is shown in Figure 3.6.

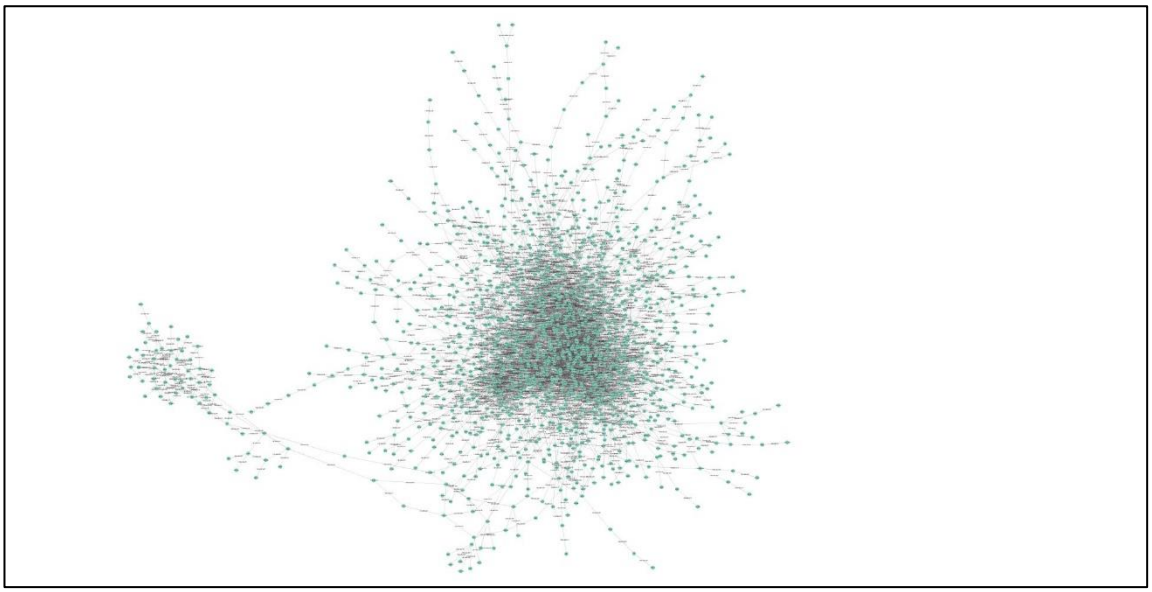

Figure 3.8: Co-Expression Network analysis for Dataset I. Decreasing correlation pattern created for Aged to Middle aged to Young groups.

3.6.3. Functional Analysis of Correlation Networks for Dataset I

Functional analysis was performed for Young to Aged decreasing correlation pattern for Dataset I. For these decreasing correlation pattern networks, functional analysis were done with g:Profiler correction method and p-value threshold was chosen as 0.05. Only the learning, memory, nervous system related terms are given in the results. This analysis does not result in the terms we are interested in such as; Learning, Memory and Plasticity, but there are other terms about neuron and synapse. Interestingly, there is a term called “long term depression” in this analysis. Moreover, Aged to Young decreasing patterns have more related terms than Young to Aged. This module includes Learning, Memory, Plasticity and also brain development and inflammatory processes. It was expected that the Learning and memory performance decrease from young to aged, but the results interestingly show the opposite. (Table 3.12, Table 3.13)

Table 3.12: g:Profiler results for Aged to Young decreasing correlation pattern for GSE854 dataset.

term type	term name	p-value
BP	regulation of signaling	3.22E-24
BP	cell-cell signaling	1.01E-19
BP	synaptic signaling	1.75E-15
BP	trans-synaptic signaling	1.75E-15
BP	nervous system development	3.56E-15
BP	generation of neurons	1.42E-10
BP	Neurogenesis	2.79E-10
BP	central nervous system development	4.09E-10
BP	regulation of synaptic plasticity	2.15E-08
BP	neuron differentiation	2.31E-08
BP	regulation of neurogenesis	4.09E-08
BP	regulation of nervous system development	4.74E-08
BP	brain development	2.99E-07
BP	negative regulation of nervous system development	2.63E-06
BP	regulation of immune system process	2.66E-06
BP	negative regulation of neurogenesis	1.05E-05
BP	regulation of cytokine production	1.13E-05
BP	neuron development	1.27E-05
BP	cytokine production	2.09E-05
BP	response to oxidative stress	2.43E-05
BP	inflammatory response	0.000101
BP	regulation of neuronal synaptic plasticity	0.000139
BP	cellular response to cytokine stimulus	0.000382
BP	cell morphogenesis involved in neuron differentiation	0.000467
BP	immune system development	0.000566
BP	learning or memory	0.00153
BP	regulation of neuron differentiation	0.00155
BP	Cognition	0.00252
BP	Gliogenesis	0.00279
BP	immune system process	0.00355
BP	Learning	0.0169
CC	neuron part	4.78E-10
CC	neuron projection	3.84E-07
keg	neuroactive ligand-receptor interaction	0.00459
MF	neurotransmitter receptor activity	0.000041
rea	neuronal System	3.83E-07

Table 3.13: g:Profiler results for Young to Aged decreasing correlation pattern for GSE854 dataset.

term type	term name	p-value
BP	cell death	0.00228
BP	programmed cell death	0.00216
BP	neuron apoptotic process	0.0279
BP	regulation of neuron apoptotic process	0.0134
BP	nervous system development	0.00419
CC	neuron part	0.000703
CC	neuron projection	0.0271
CC	Synapse	0.00785
CC	synapse part	0.0277
MF	ion binding	0.0301
keg	neurotrophin signaling pathway	0.00215
keg	long-term depression	0.00603

3.7. Weighted Gene Correlation Network Analysis for Dataset I

3.7.1. Detection and Removal of Outliers

GSE854 [Blalock et al. 2003] has 3 age groups, Young, Middle aged and Aged. Each age group has 10 samples. For Weighted Gene Correlation Network Analysis (WGCNA), the normalized transcriptome data was given as an input to the algorithm. WGCNA has its own sample clustering algorithm, which uses hierarchical clustering. “hclust” function was used to create clustering tree with method= “average”. This algorithm allows user to define a threshold to cluster samples. Clustering creates a tree and each branch in this tree represents one sample (Figure 3.9). It is assumed that if there is an outlier in this data that outlier is located far from its group. The dissimilarity value for clustering was chosen as 20 and this threshold led to the identification of four samples as outliers. These are; Aged samples 3, 4 and 8 and Young sample 4. Middle aged 8 was already defined as outlier in PCA analysis previously. (Figure 3.3, 3.4). Further analyses were performed without those four samples.

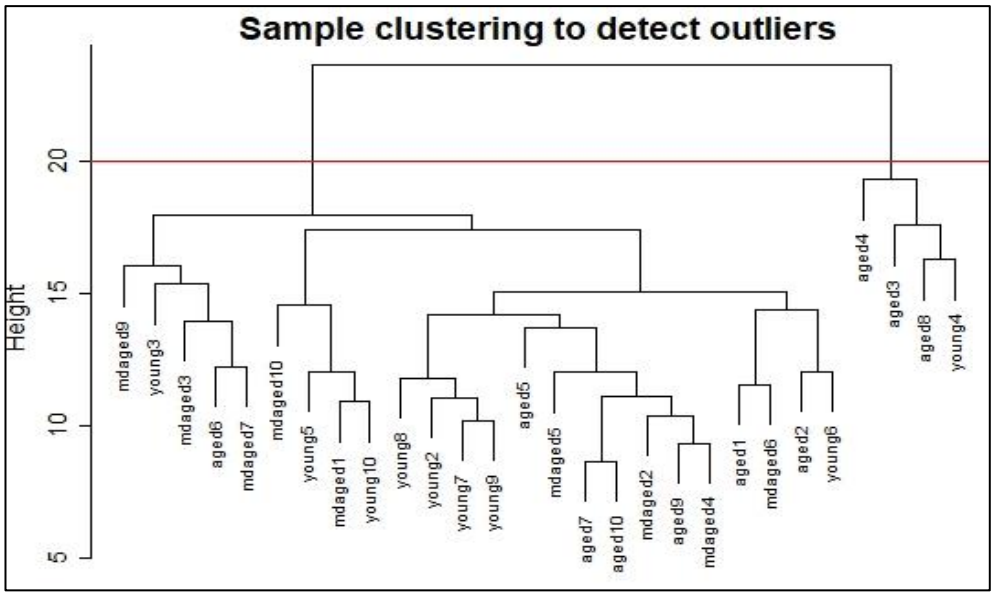

Figure 3.9: Sample clustering tree for GSE854 dataset. This analysis identified four samples as outliers.

3.7.2. Correlating Transcriptome Samples with Phenotypic Trait Data

In WGCNA package, the relationship between the trait values and the transcriptome samples can be detected. Trait data is a numeric value that includes SWM and OMT test results and that can be used as an extra information while important modules are detected. Dataset I (GSE854) has two types of trait data: OMT (Object Memory Task) and SWM (Morris Water Maze) traits. For each sample (rat), there are OMT and SWM scores based on how they performed in those tests. There are numeric data, and they can be incorporated into the WGCNA as a "trait data". Moreover, defining the three age groups as a trait data would be beneficial to label samples of each group with a different color-code. This trait was called "age" trait. "To create a numerical trait vector for the "age" trait, each sample in the same age group was assigned the same arbitrary number. For example, 9 samples of the Young group were assigned "100", and 7 samples of the Aged group was assigned "1". Also, we created New trait data and added as a matrix to this trait data to make modules condition specific (Table 3.14). In this situation each age group defined as a trait condition, and introduced as a matrix, and in this matrix each only defined condition have "1", others get "0". For example, if trait condition is "aged", only Aged samples gets "1", other age groups gets "0". Same procedures applied for three age group. And

this application based on information from website which was belong to one of the creators of WGCNA package in Web 5.

The trait data was introduced to the package, and sample dendogram was plotted together with trait data, visualized as a heatmap (Figure 3.11). The plot shows the relationship between samples and their memory related test performance. In the heatmap, darker color shows a higher score for that sample, while lighter color represents lower scores. Young samples show better performance than Aged samples based on their SWM and OMT scores. Time spent to find platform (SWM) is low, and time to analyse new object (OMT) is high for Young samples, and this is shown in dark red in the heatmap graph under the sample dendrogram (Figure 3.11). White color shows that the corresponding score for that sample is low compared to other samples. "plotDendroAndColors" and "hclust" functions are used to create these graphs.

Table 3.14: Trait data for GSE854 dataset. This table consists of scores of SWM and OMT traits from the corresponding article.

Animal ID	GSM accession	SWM Task	OMT Task	AgingTrait	Aged	Middle Aged	Young
aged1	GSM13303	0	0.073	1	1	0	0
aged2	GSM13304	2	-0.138	1	1	0	0
aged3	GSM13305	0	0.182	1	1	0	0
aged4	GSM13306	3	0.261	1	1	0	0
aged5	GSM13307	0	-0.529	1	1	0	0
aged6	GSM13308	0	0.231	1	1	0	0
aged7	GSM13309	1	-0.448	1	1	0	0
aged8	GSM13310	3	-0.269	1	1	0	0
aged9	GSM13311	1	0.069	1	1	0	0
aged10	GSM13312	2	-0.169	1	1	0	0
mdaged1	GSM13313	4	0.333	50	0	1	0
mdaged2	GSM13314	3	0.091	50	0	1	0
mdaged3	GSM13315	4	0.519	50	0	1	0
mdaged4	GSM13316	0	0.048	50	0	1	0
mdaged5	GSM13317	5	0.831	50	0	1	0
mdaged6	GSM13318	2	-0.084	50	0	1	0

Table 3.14: Trait data for GSE854 dataset. This table consists of scores of SWM and OMT traits from the corresponding article (continue).

Animal ID	GSM accession	SWM Task	OMT Task	AgingTrait	Aged	Middle Aged	Young
mdaged7	GSM13319	2	0.273	50	0	1	0
mdaged9	GSM13321	2	0.373	50	0	1	0
mdaged10	GSM13322	3	0.429	50	0	1	0
young2	GSM13323	2	0.081	100	0	0	1
young3	GSM13324	3	0.219	100	0	0	1
young4	GSM13325	3	0.421	100	0	0	1
young5	GSM13330	2	0.67	100	0	0	1
young6	GSM13326	4	0.375	100	0	0	1
young7	GSM13327	5	0.44	100	0	0	1
young8	GSM13328	2	0.451	100	0	0	1
young9	GSM13329	5	0.43	100	0	0	1
young10	GSM13331	2	0.731	100	0	0	1

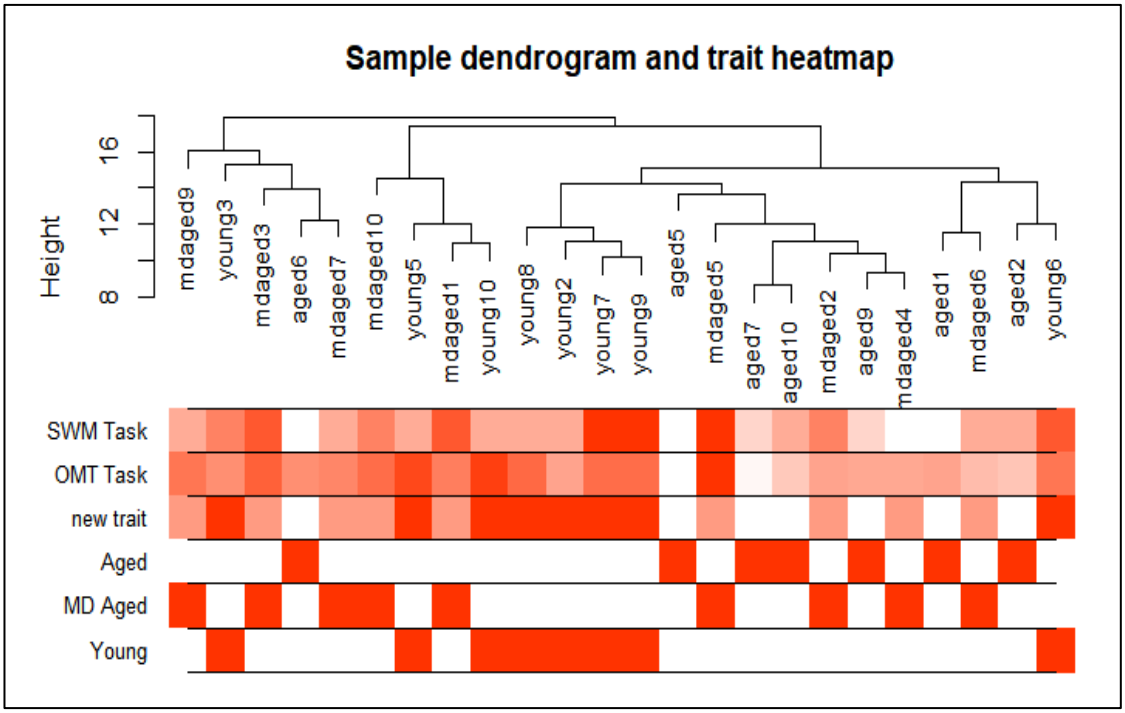

Figure 3.11: Sample dendrogram and trait heatmap plot for GSE854 dataset. This figure shows the relationship between the transcriptomic samples and the traits.

3.7.3. Soft-Threshold Determination

The next step in WGCNA package is choosing a soft threshold (power) that can make the output co-expression network scale free. Scale Free Topology means most of the nodes in the network has low number of connectivity scores, and there are a few nodes in the network with very high connectivity scores. The high-connectivity nodes are called hubs. Since biological networks are known to show scale free topology, WGCNA aims to obtain a correlation network that is scale-free in terms of network structure. Considering expression data, the range of power was created from 0 to 20. *pickSoftThreshold* function creates a figure that shows scale freeness of the output network for each soft threshold value from 0 to 20. Soft-threshold value can be chosen at the point where scale freeness gets a stable value. Each value on the figure shows the "power" value that makes co-expression network scale free by raising the power of each correlation value to the chosen parameter. The soft threshold was chosen as 12, which corresponded to 80% scale free rate (Figure 3.12). The line gets straight at that point and increase in scale freeness is minimum for further points (in this case, cor=cor^{12}).

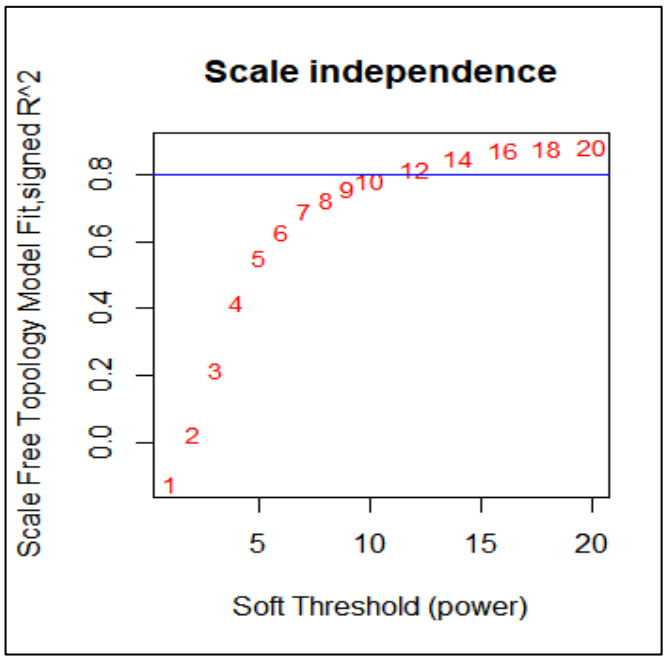

Figure 3.12: Determination of the power parameter for GSE854 dataset. The model was fitted at 12 to make co-expression matrix scale free in 80% percentage.

3.7.4. Module Creation

The next step is creating modules by using expression data and power. Modules are created by *blockwiseModules* function. This function requires expression data and the power parameter. There are some other parameters important for this analysis; networkType was changed to "signed". This parameter shows the sign of the correlation, negative or positive, and if this parameter turned into "signed" that means only the positive correlations are considered for further analysis. A positive correlation means gene pairs show similar expression profiles, and they belong to same functional mechanism. However, negative correlation shows expression changes in opposite directions. Negatively correlated genes have lower chance to belong to the same functional mechanism compared to the positively correlated ones. "minModuleSize" was adjusted to 40, this parameter defines minimum allowed number of the nodes in the modules. 40 means that at least 40 genes will hbe placed into the modules. mergeCutHeight was adjusted to 0.25, and this parameter specifies the height at which modules are combined. Other parameters were set to default. Modules were created with those parameters. There were 25 modules created and 89 genes do not belong to any module. WGCNA assigns each module to a color. If a node does not belong to any module it goes to grey module (Figure 3.13). Grey module is defined as dysfunctional module.

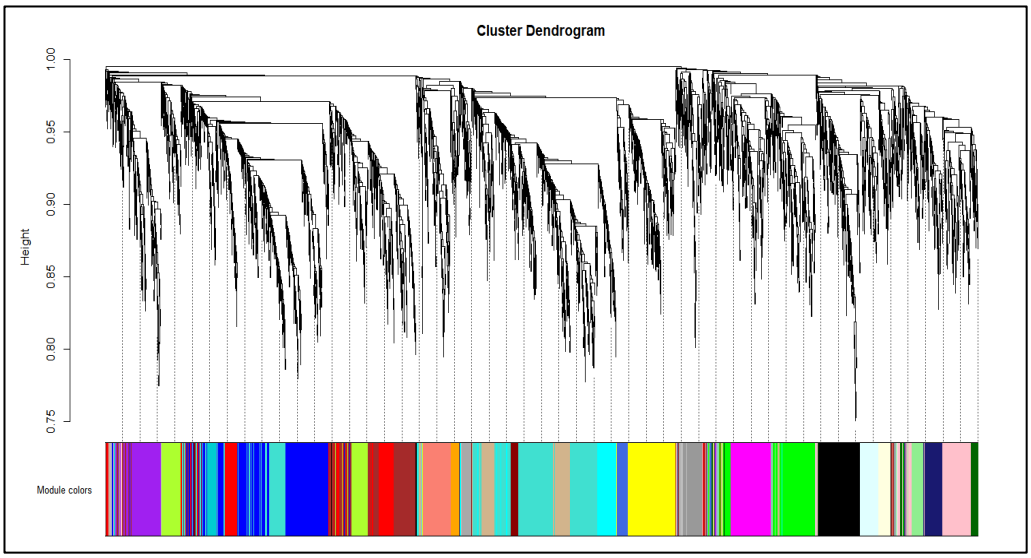

Figure 3.13: Cluster dendrogram for GSE854 dataset. Modules are created and assigned to a single color. There were 25 modules and each of them had 40 nodes at least.

Table 3.15: Color of modules and their number of nodes.

Color of module	Number of nodes	Color of module	Number of nodes	Color of module	Number of nodes
Black	255	Greenyellow	211	Purple	219
Blue	443	Grey60	109	Red	259
Brown	331	Lightcyan	116	Royalblue	74
Cyan	133	Lightgreen	90	Salmon	163
Darkgreen	62	Lightyellow	78	Tan	210
Darkgrey	51	Magenta	224	Turquoise	687
Darkred	66	Midnightblue	117	Yellow	274
Darkturquoise	54	Orange	46	Grey	89
Green	274	Pink	238		

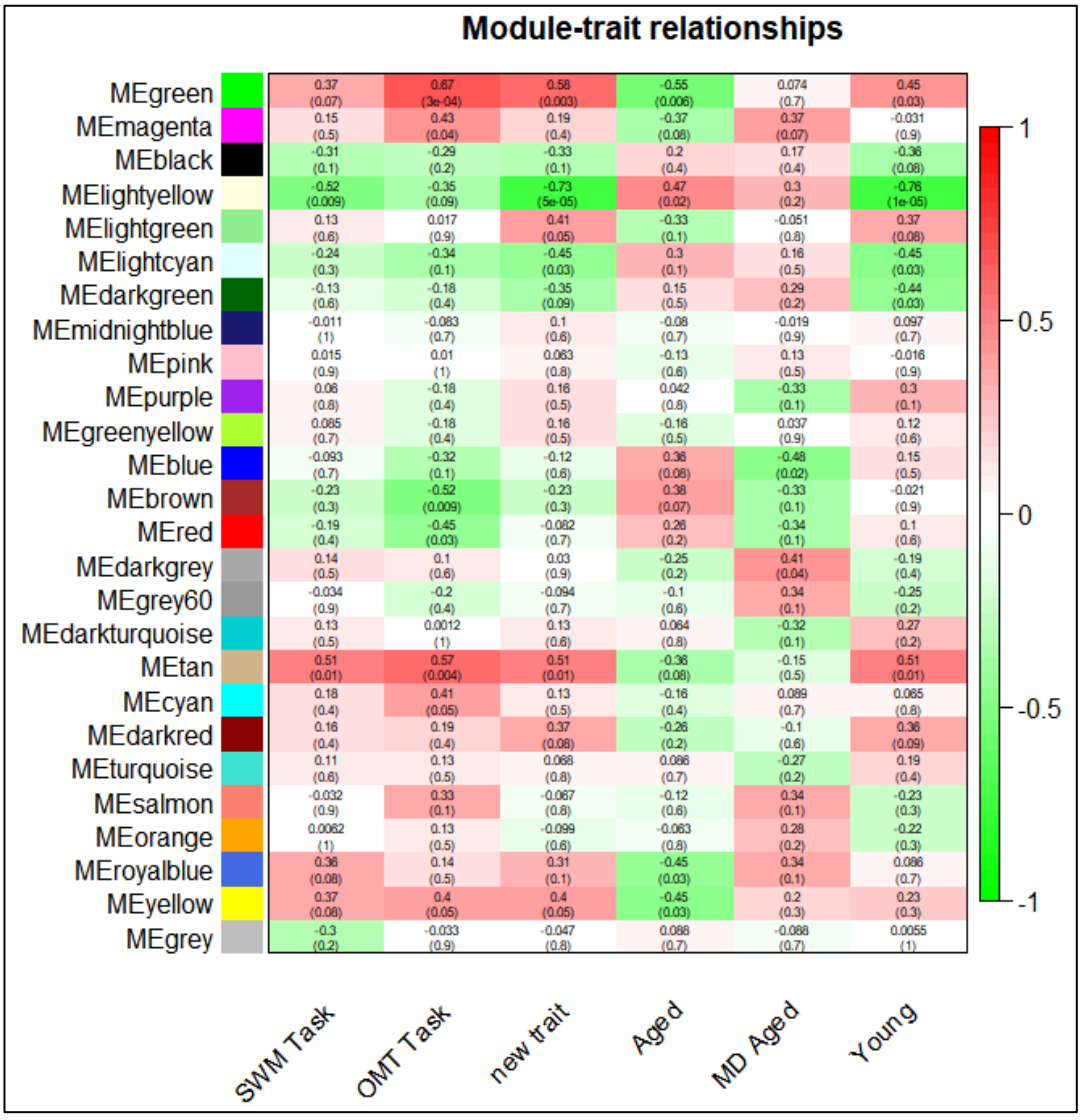

Figure 3.14: Module-trait relationship for Dataset I. This figure shows the relationship of modules to each trait condition. Values in parantheses show p-values.

In WGCNA, each module is represented by a color and genes in the modules are highly correlated with each other or they have similar expression profiles. This algorithm calculates eigengene vector for each module, which is created by using correlation values of genes and it has as many elements as the number of samples. Module eigengene is defined as the first principle component of a module. It can also be defined as the gene expression representation of a module [Langfelder et al., 2008]. Eigengene is created by using expression profile of genes in the module (Figure 3.15). Eigengene is specific for the module and this vector is mathematical representation of a module. It is used to find significant modules with respect to trait data. Trait data is the numerical data which includes SWM and/or OMT performance of samples introduced to the WGCNA. The correlation is calculated between trait data and eigengene vector, if they have higher correlation with each other, genes or processes in that module significantly change for that trait condition [Langfelder 2007].

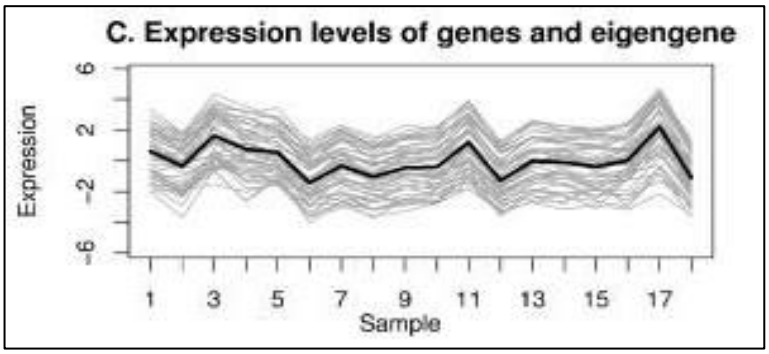

Figure 3.15: Creation of module eigengene

Module trait relationship graph can be used to find important modules among dozens of modules. Module trait relationship graph can be created by the package by using trait data. Modules associated with different traits can be detected by using this graph (Figure 3.14). Tan and green modules show high correlation with four trait conditions, lightyellow shows high correlation for Aged trait condition and darkgrey module is important for Middle-aged trait condition.

Module-Membership (MM) versus Gene Significance (GS) plot can also be created by the package for important modules in trait conditions. Module Membership shows how correlated gene “i” is with the eigengene of module “x”. MM shows the correlation of each gene with the eigengene of a module to understand the relationship of gene “i” with the given module. If this value is close to 0, this gene is not part of

the related module, if it is close to "1" this gene is highly correlated with the genes of related module. Gene significance shows the biological significance of gene "i" in related modules. This represents the relationship of gene "i" with the module based on external information, like SWM and OMT performance. In other words, GS shows the significance of gene "i" in a module based on the physical performance. In this study trait data was used. Gene Significance versus Module Membership graph is created for the chosen trait to see gene "i" in given module and the distribution of genes in that module. Gene significance is used to add external information to co-expression network [Langfelder et al., 2008]. These graphs were created for tan and green modules because of their higher correlation with SWM and OMT trait conditions. (Figure 3.16, 3.17). In Tan module, the correlation of Module Membership of genes in the module and significance of each gene for OMT trait condition is 0.48. In green module, these values are more significant, meaning that at OMT trait condition genes are more related to each other and with the module.

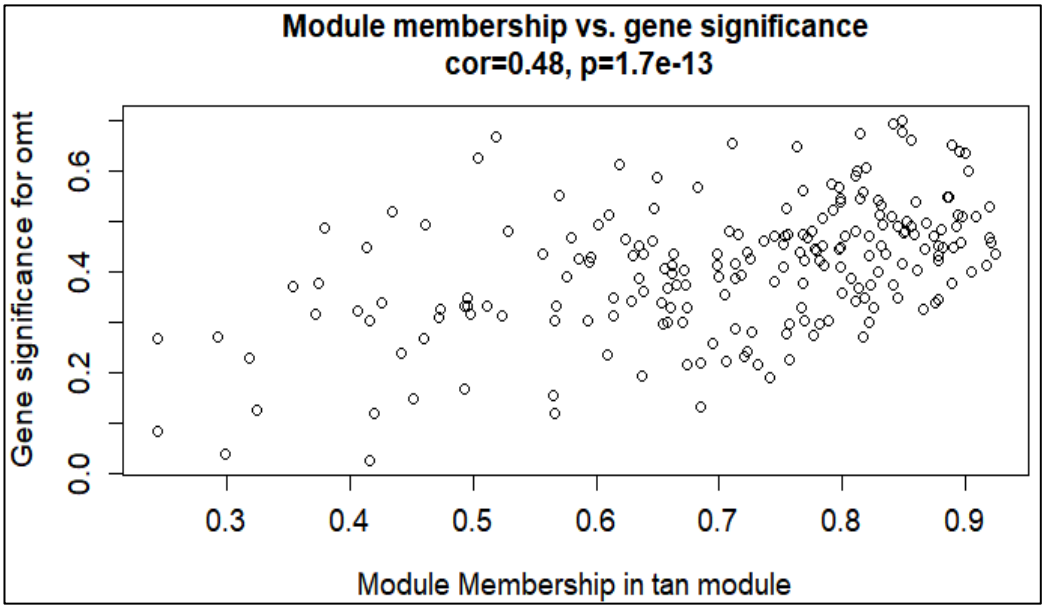

Figure 3.16: GS versus MM graph for tan module.

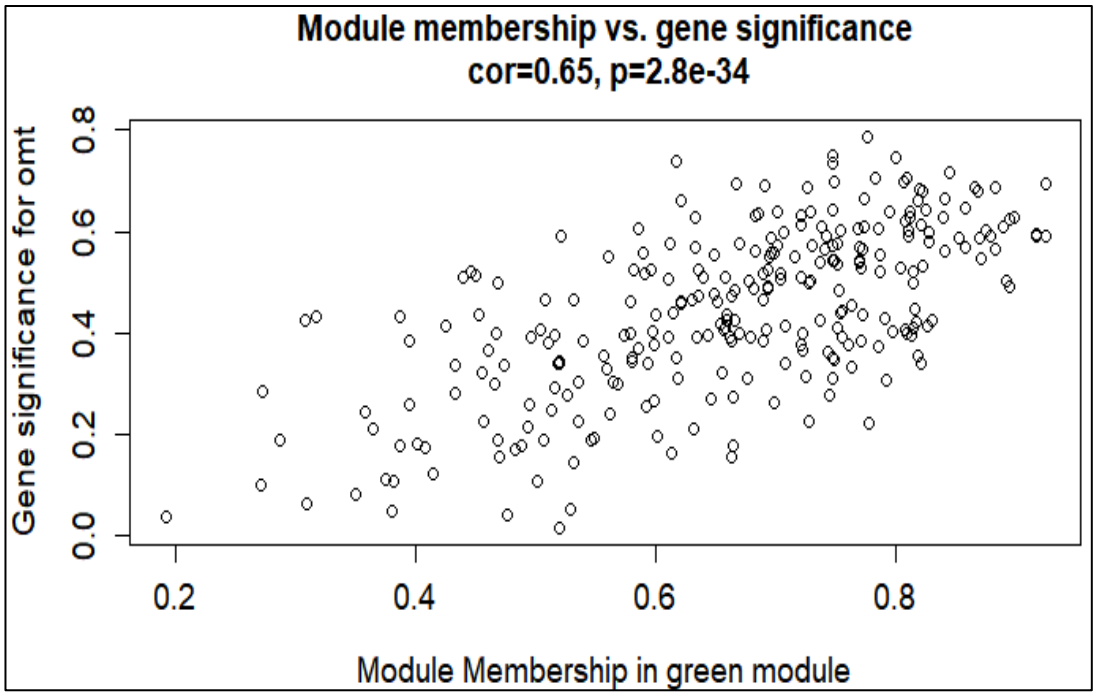

Figure 3.17: GS versus MM graph for green module.

3.7.5. Functional Analysis of Results by using UserListEnrichment function

To check the functional ingredients of modules, *userListEnrichment* function was used in WGCNA package. *userListEnrichment* function can be used if there are gene lists that belong to specific functional units, and we want to see if genes in modules are enriched with these gene lists. This function also has a parameter "useBrainList". When this parameter is activated a pre-defined gene list that included genes related to functions about brain, neuron, plasticity etc is used as input to the function. We can use these pre-defined lists, or we can create our own lists for this function.

For this study, the gene lists were created by using Gene Ontology (GO) [The Gene Ontology Consortium, 2019]. Five different gene lists were created using GO, these lists are; synaptic plasticity, learning and memory, neurogenesis and gliogenesis, cognition, and brain, which includes all lists. These terms are searched in GO web page and with "Annotation" section organism can be chosen. The lists include gene names, which are related to the that function. In Table 3.16, the sizes of the lists are given.

Table 3.16 Gene Lists. These lists were created by searching GO.

List Name	Gliogenesis and neurogenesis	Learning and Memory	Synaptic Plasticiy	Cognition	Brain (Total)
Number of genes	1823	327	1220	350	2740

UserListEnrichment function has four inputs; (i) “geneR” includes all gene symbols that are sorted by their module color (ii) ”labelR” includes module colors in the same order as the gene symbols, (iii) “fnln” includes the user-defined or pre-defined gene lists in text file whose enrichment in the modules will be tested (iv) ”catNmln” defines the names of the lists. This function has also its own pre-defined lists about brain, we can use these lists by turning the related parameter to “TRUE” (useBrainList =” TRUE). The pre-defined Brain lists were not used in this study. All lists were used together and based on the results important modules were detected (Table 3.17) and their functional analysis were done. This table (Table 3.17) gives the result of UserListEnrichment analysis. The first column of the table defines the modules that were enriched in terms of the functional categories given in Table 3.16. The category, to which the module is belong, is also given in the table. At the right side of the table, significance of the association of a module with the corresponding user defined category is given. Black, Salmon and Ligthcyan modules are significantly enriched with synaptic plasticity and gliogenesis and neurogenesis terms. They include genes that commonly function in memory and learning related processes. Therefore, these modules were further analyzed with g:Profiler. Visualization of modules was performed in Cytoscape. Salmon module is shown in Figure 3.18.

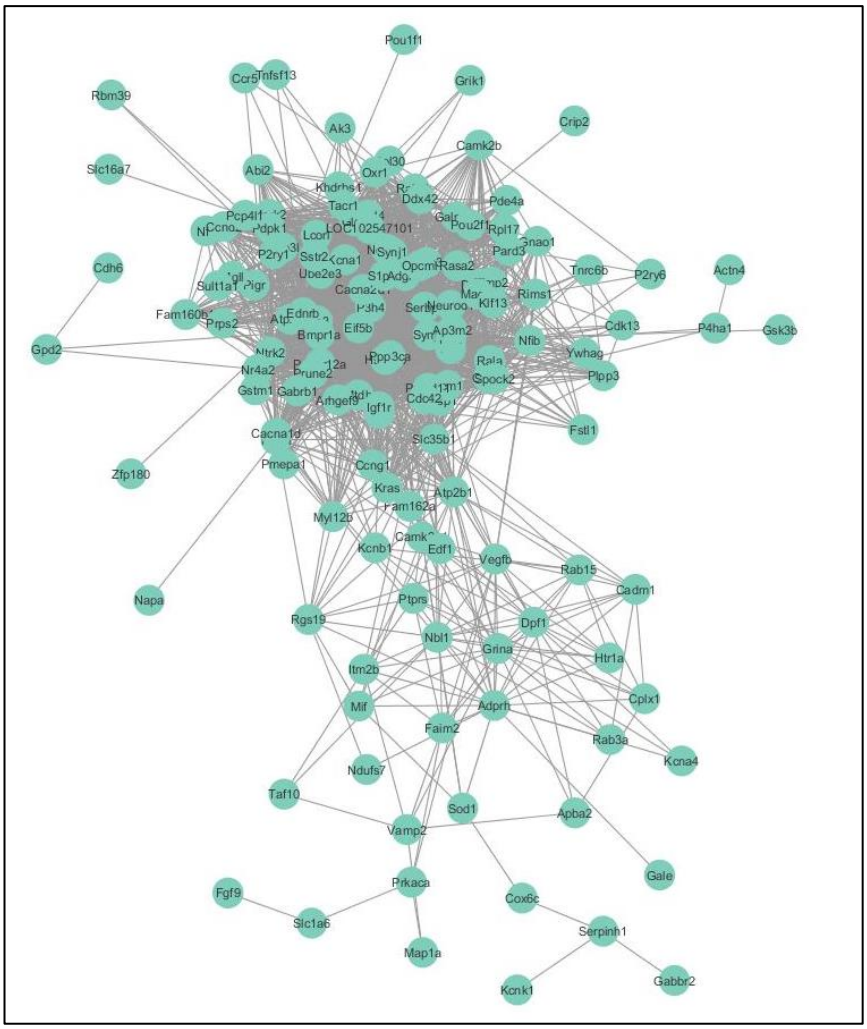

Figure 3.18 WGCNA Salmon module for Dataset I.

Table 3.17: *userListEnrichment* function analysis result. This table contains color and number of genes information for important modules based on their significance (Corrected p-value). Corrected p-value gives information about significance of modules for that functional category. These lists were created for this study and they are related to brain, learning and memory.

Module color	User Defined Categories	Corrected p-values
Black	Brain	1.06E-06
salmon	Synaptic plasticity	3.37E-06
salmon	Brain	1.43E-05
Black	Synaptic plasticity	2.37E-05
lightcyan	Brain	4.11E-05
Black	Gliogenesis and neurogenesis	0.000112
lightcyan	Synaptic plasticity	0.000199
lightcyan	Gliogenesis and neurogenesis	0.006009

3.7.6. Functional analysis of WGCNA modules for Dataset I

Black, Salmon and Ligthcyan modules are found to be important as a result of UserListEnrichment-based analysis, they have low significant p-values. These modules are significantly associated with memory and learning related terms. These modules were analysed with g:Profiler correction method with 0.05 threshold, and only memory, learning, nervous sytem related results are given in Table 3.18-21. Black module has aging related cell death and apoptotic processes. Learning, Memory terms are not identified in this analysis but there are other related terms such as Nervous System, neuron, signaling. Aging is detected in Salmon module and this module is one of the best modules for WGCNA analysis in terms of memory related function. Lightcyan is the second-best module and it includes cognition and plasticity terms. Pink module is also analysed with g:Profiler and it is the only module with Learning and Memory terms. This module is the leeast significant module in terms of correlation of module with trait data.

Table 3.18: g:Profiler analysis for Salmon module.

term type	term name	p-value
BP	neuron apoptotic process	0.00248
BP	regulation of neuron apoptotic process	0.00732
BP	neurotransmitter transport	0.000961
BP	Signaling	0.0083
BP	Aging	0.00257
BP	nervous system development	5.55E-09
BP	central nervous system development	7.91E-08
BP	brain development	4.16E-08
BP	forebrain development	0.000285
BP	Neurogenesis	6.6E-07
BP	generation of neurons	0.000024
BP	neuron differentiation	0.00149
BP	cell-cell signaling	6.65E-08
BP	synaptic signaling	0.00101
BP	trans-synaptic signaling	0.00101
BP	regulation of nervous system development	0.00661
BP	regulation of signaling	4.25E-06
BP	negative regulation of signaling	0.0133
BP	neurotransmitter secretion	0.00255
CC	neuron part	1.25E-11
CC	neuronal cell body	0.00143
CC	neuron projection	1.31E-09
CC	neuron projection terminus	0.0249
Keg	Neuroactive ligand-receptor interaction	0.000818
Rea	Serotonin Neurotransmitter Release Cycle	0.0132
Rea	Dopamine Neurotransmitter Release Cycle	0.0434
Rea	Norepinephrine Neurotransmitter Release Cycle	0.0166
Rea	Glutamate Neurotransmitter Release Cycle	0.00237
Rea	Neuronal System	2.11E-05

Table 3.19: g:Profiler analysis for Lightcyan module.

term type	term name	p-value
BP	Cognition	0.0227
BP	neurotransmitter transport	0.027
BP	nervous system development	0.000218
BP	regulation of nervous system development	0.0487
BP	positive regulation of nervous system development	0.00413
BP	regulation of neurotransmitter levels	0.0182
BP	cell communication	0.00484
BP	cell-cell signaling	7.35E-05
BP	synaptic signaling	4.84E-06
BP	trans-synaptic signaling	4.84E-06
BP	neurotransmitter secretion	0.00266
BP	regulation of synaptic plasticity	0.049
BP	Neurogenesis	0.00243
BP	generation of neurons	0.000645
BP	neuron differentiation	0.0247
BP	neuron development	0.00437
BP	positive regulation of neurogenesis	0.00111
BP	positive regulation of neuron differentiation	0.0045
BP	neuron projection development	0.00327
BP	neuron projection morphogenesis	0.0011
BP	regulation of neuron projection development	0.0492
BP	positive regulation of neuron projection development	0.0213
CC	neuron part	5.21E-10
CC	neuron to neuron synapse	1.76E-05
CC	neuron projection	2.92E-06
CC	neuronal cell body	0.00297

Table 3.20: g:Profiler analysis of Black module.

term type	term name	p-value
BP	cell death	0.00228
BP	programmed cell death	0.00216
BP	neuron apoptotic process	0.0279
BP	regulation of neuron apoptotic process	0.0134
BP	nervous system development	0.00419
CC	neuron part	0.000703
CC	neuron projection	0.0271
CC	Synapse	0.00785
MF	ion binding	0.0301
Keg	Neurotrophin signaling pathway	0.00215
Keg	Long-term depression	0.00603

Table 3.21: g:Profiler analysis for Pink module.

term type	term name	p-value
BP	synaptic signaling	0.0157
BP	nervous system development	0.000503
BP	regulation of nervous system development	0.029
BP	Neurogenesis	0.000798
BP	generation of neurons	0.000344
BP	neuron differentiation	0.000203
BP	neuron development	3.87E-05
BP	neuron projection development	9.99E-05
BP	neuron projection morphogenesis	0.000687
BP	cell morphogenesis involved in neuron differentiation	0.00262
BP	learning or memory	0.0366
CC	neuron part	2.04E-06
CC	neuron projection	1.98E-05

4. NETWORK BASED ANALYSIS OF MEMORY-ASSOCIATED TRANSCRIPTOME: DATASET II

4.1. PPI for *Rattus Norvegicus*

In the analysis of Dataset II, the same PPI data called RatMusInt, which was created in Chapter 3.1, was used for subnetwork discovery. It has 9500 genes (nodes) and 37043 interactions (edges).

4.2. Dataset Description

The second dataset is GSE5666 [Rowe et al., 2007], and it uses hippocampal gene expression analysis to find cognition related pathways via aging. *Rattus norvegicus* was used as the model organism in the study. There were 78 samples in total and there were three categories;

- Non-Trained animals (NT), which are divided into two sub-groups and have 10 samples for each; Young (Y) and Aged (A). These animals did not have any training. Only the effect of aging can be seen by comparing them. Young animals are 3-4 months old and Aged animals are 23-24 months old.
- 5-Day Training group (5T), which has three sub-groups; Young (Y), Aged Unimpaired (AU) and Aged Impaired (AI). Each sub-group has 10 samples. These animals had training for five days and at the end of the training (fifth day) the animals are sacrificed, and transcriptome data is collected from the hippocampal samples of their brain. Also, Aged animals were divided into two categories, impaired and unimpaired, based on their performance in memory performance tests in comparison to young samples.
- 21-Day Post-training animals, also having three sub-groups; Young (Y), Aged Unimpaired (AU) and Aged Impaired (AI) and have 10 samples for each. These animals had training and after 21 days after the end of the 5-day training the animals were sacrificed and transcriptome data was collected from the hippocampal samples of their brain. The overall design of the experiments is shown in Table 4.1.

Table 4.1: The overview of experimental design for GSE5666 dataset.

Training days/ Animals	5T (5 day of training and immediately sacrified)	21PT (Sacrified 21 days post training)	NT (No training)
Young (Y)	5T-Y N=10	21PT-Y N=9	NT-Y N=10
Aged Umpaired (AU)	5T_AU N=10	21PT-AU N=10	NT-A N=10
Aged Impaired (AI)	5T-AI N=10	2PT-AI N=9	

Morris Water Maze test are applied to the rats in this study. Based on their performance Aged samples are divided into two groups, which are Aged Unimpaired (AU) and Aged Impaired (AI). The performance of the Young and Aged samples should be similar but not the same. Latency is most important result of this test and shows the time spend to find platform that is placed into the water maze. During training, latency should decrease. If the performance of aged sample is noticably different compared to Young sample and to other aged samples, it is grouped under Aged Impaired group. In Aged Impaired animals, latency does not decrease with more training, it shows zigzag pattern. Not only the age affects the performance of these samples, but also stress, physical deficiencies or cognitive disorders may influence their performance. Aged Unimpaired animals are assumed to have only aging effect on their performance when compaired to the Young samples.

In the original study, two different statistical tests were applied to the different training groups [Rowe et al., 2007]. First Non-Trained animals were analyzed with Student t-test and differentially expressed genes were identified. And, ANOVA was used to find differentially expressed genes between 5-Day Training and 21-Day Post Training samples. Down-regulated genes were identified, and GO analysis was performed for these genes. Aging and cognitive related functions were identified.

4.3. Preprocessing of Dataset

4.3.1. Dataset Normalization

Dataset II consists of 78 samples. This data was downloaded from Gene Expression Omnibus (GEO) database (Clough et al 2016) with the accession number GSE5666. As explained in Chapter 3.3, raw microarray data was downloaded from GEO and normalized using RMA package.

4.3.2. PCA/Sammon mapping to Detect Outliers

GSE5666 data consisted of 78 samples in total. To see the distribution of these samples with respect to each other in terms of their gene expression values and to find if there were any outliers, we performed PCA analysis and Sammon mapping in MATLAB. With PCA analysis, we can see the distribution of samples even when samples from many different groups are considered, but in Sammon mapping only two groups of samples can be compared. PCA and Sammon mapping were applied to this data in different combinations. First, all 78 samples were used in PCA analysis, then PCA was applied to 5T and 21PT training groups together, afterwards 5T and 21PT groups were separately analyzed with PCA, and finally PCA was applied to NTY and NTA groups together. Sammon mapping was done for two groups; first group consists of 5T and 21PT samples and second one consists of NTY and NTA samples. Based on these analyses, 21PT-Y3 and 5T-AI4 were detected as outliers in both analyses. They were removed from the data for further analysis.

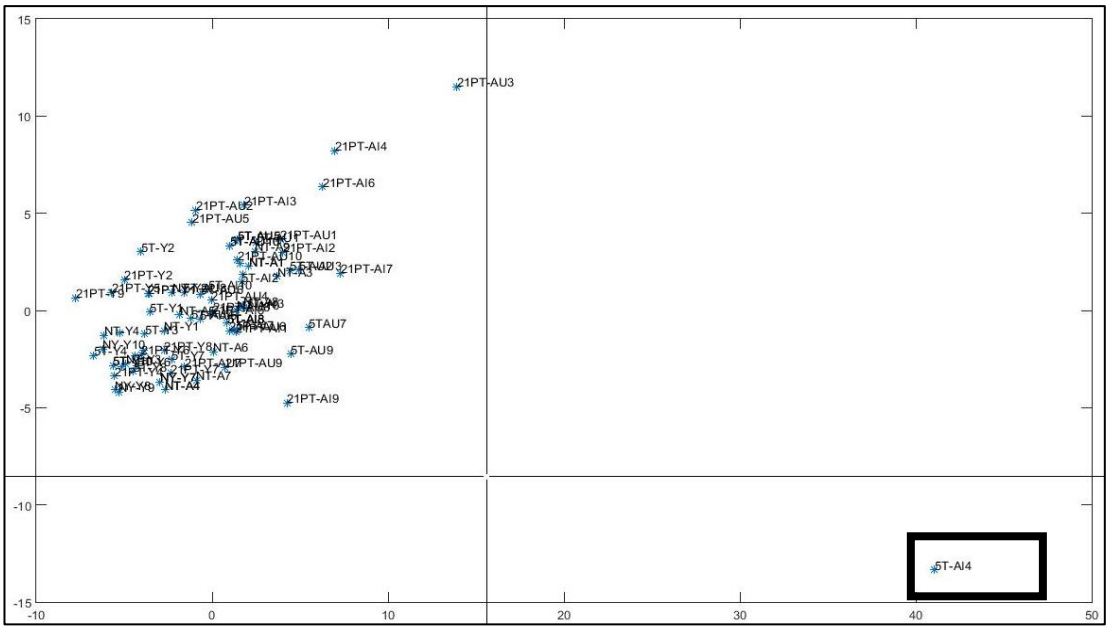

Figure 4.1: PCA analysis for GSE5666. All samples are considered for this analysis. 5T-AI4 is an outlier.

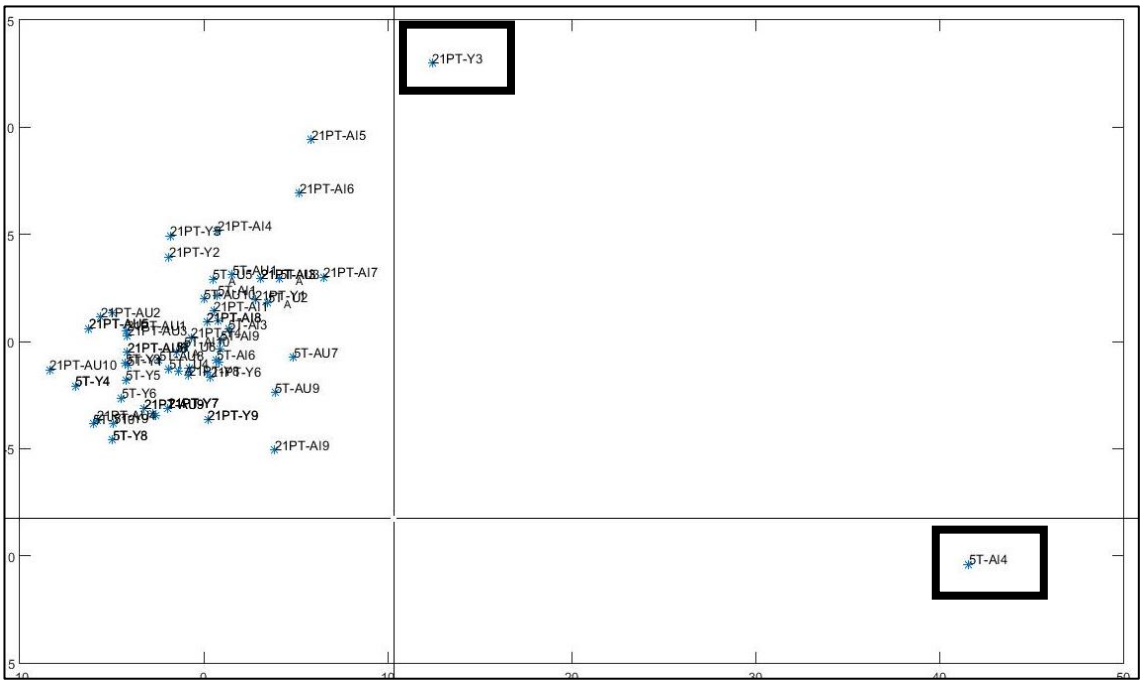

Figure 4.2: PCA analysis for GSE5666. 21PT and 5T groups are considered. 21PT-Y3 and 5T-AI4 are outliers.

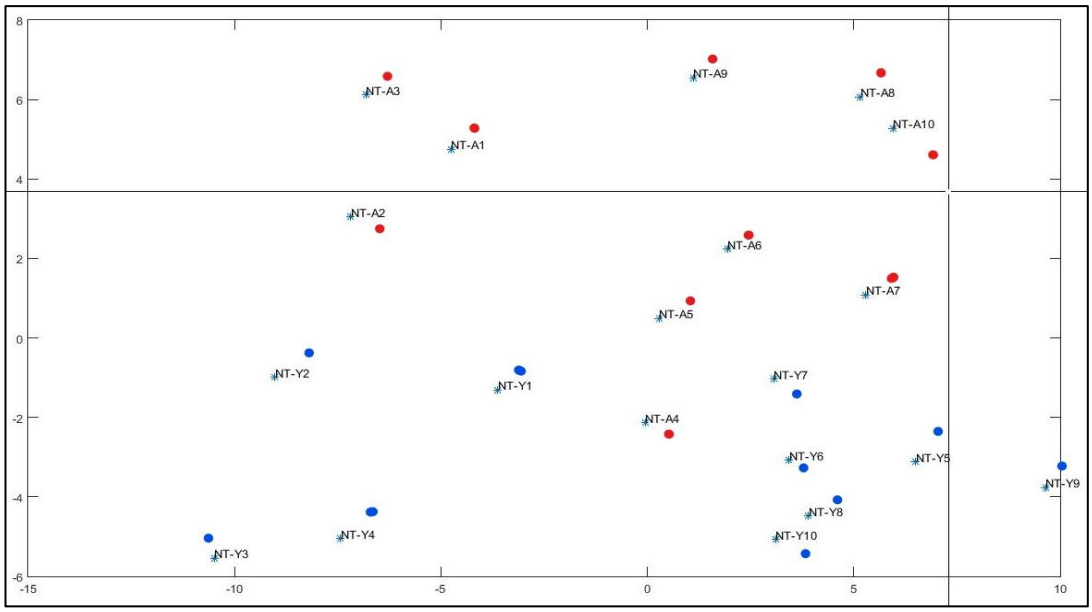

Figure 4.3: PCA analysis for GSE5666. PCA was done to the NTA and NTY samples together. There was no outlier for this condition.

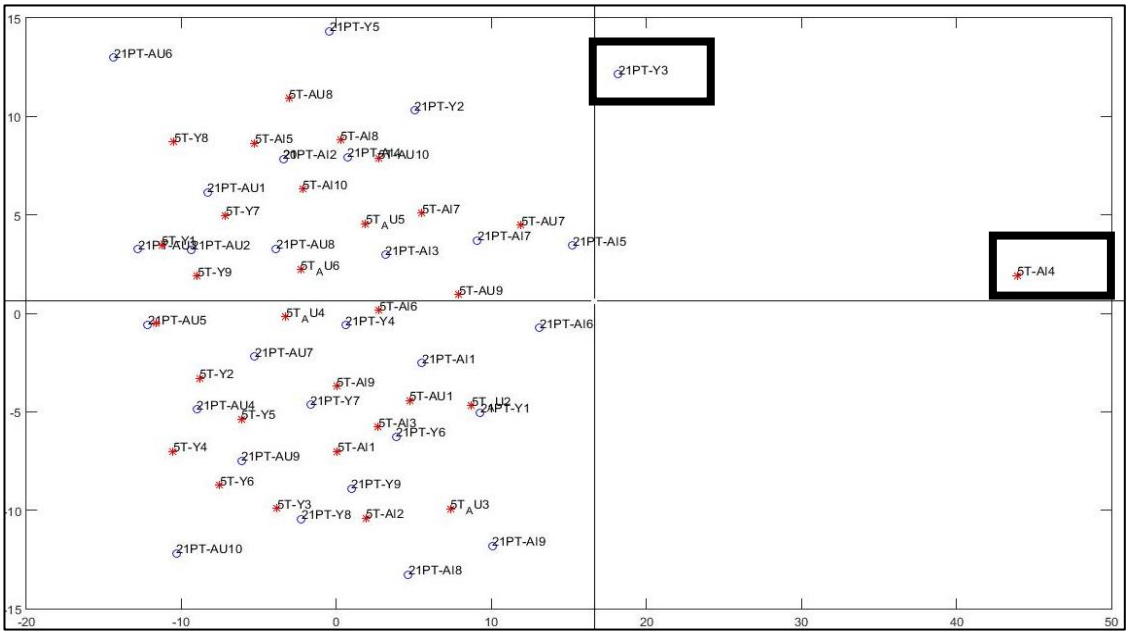

Figure 4.4: Sammon mapping for GSE5666. 21PT and 5T samples are considered. 5T-AI4 and 21PT-Y3 are outliers.

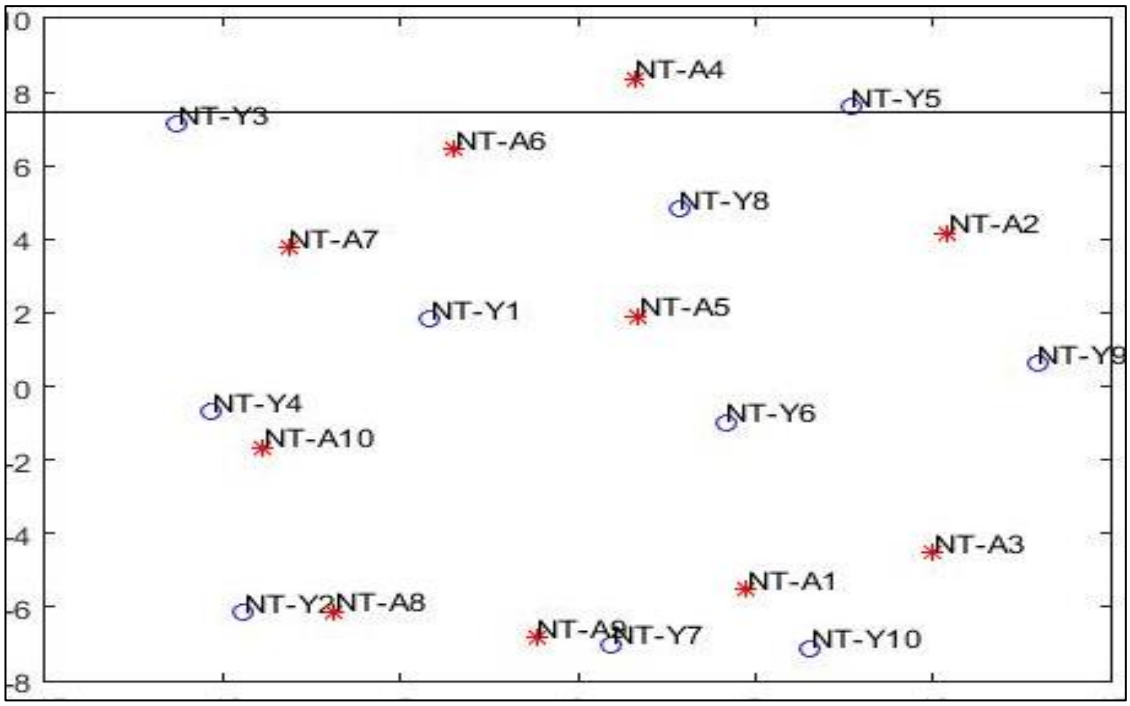

Figure 4.5: Sammon mapping for GSE5666. NTA and NTY groups are considered. There were no outliers for this condition.

4.3.3. Calculation of P-values

The third process after normalization and PCA-Sammon mapping is the calculation of p-values. The p-value of each gene is calculated and used as a score for each gene [Beisser et al. 2010]. Gene names/IDs and corresponding p-values are needed for further analysis. Normalized data was used to calculate p-values. Limma package (R) was used for this purpose.

There are three main groups in the dataset, which are given in Chapter 4.2. P-value calculation was performed for one group: 5 day Trained Young (5TY) and 5 day Trained Aged (5TAI/AU). This was specifically chosen since, in this way, the same comparison as performed in Dataset 1 (GSE854) in Chapter 3.3 was repeated with a different dataset. In Dataset 1, one of the major comparisons was between two age groups (young and old), and their samples were collected at the last day of training. The samples were also collected at the last day of training in 5T groups.

There were 15923 probes in the transcriptome data. 2349 probes did not have a gene name and they were removed from the transcriptome data. There were at least two gene symbols for 670 probes, and all symbols were included for further analyses. There were more than one probes for some genes, and we used the most significantly

changed probes to represent such genes. Thanks to this pre-processing step, all genes and their p-values were made unique. Finally, 11860 genes and their p-values were used for further analyses.

4.4. KeyPathwayMiner (KPM) Analysis for Dataset II

For KPM analysis, interactome ratmusint was introduced to Cytoscape. Preprocessed transcriptome data was also given as an input to KPM. Data was reorganized and it was converted into binary format based on the chosen threshold data. The threshold value of 0.01 resulted in 924 significantly changed genes, and this threshold was used for further analysis. Then subnetworks were identified by changing K values. 72% of the transcriptome data mapped on the interactome data, and 58% of the interactome was used. This means nearly 3506 genes were mapped on interactome. Each subnetwork was saved, and their topological analysis were performed (Table 4.2).

Table 4.2: KPM results for GSE5666 dataset. KPM was performed for three different K values.

Age group/ Explanations of data:	K=0	K=1	K=2
	Size of networks and highest degree nodes		
Young (10) – Aged (19) p<0,05=1769 **p<0,01=924**	97 node - 109 edge Irf8: 27 Stat3: 10	191 node - 223 edge Fancd2: 97 Irf8: 27	240 node - 311 edge Fancd2: 98 Ubc : 85

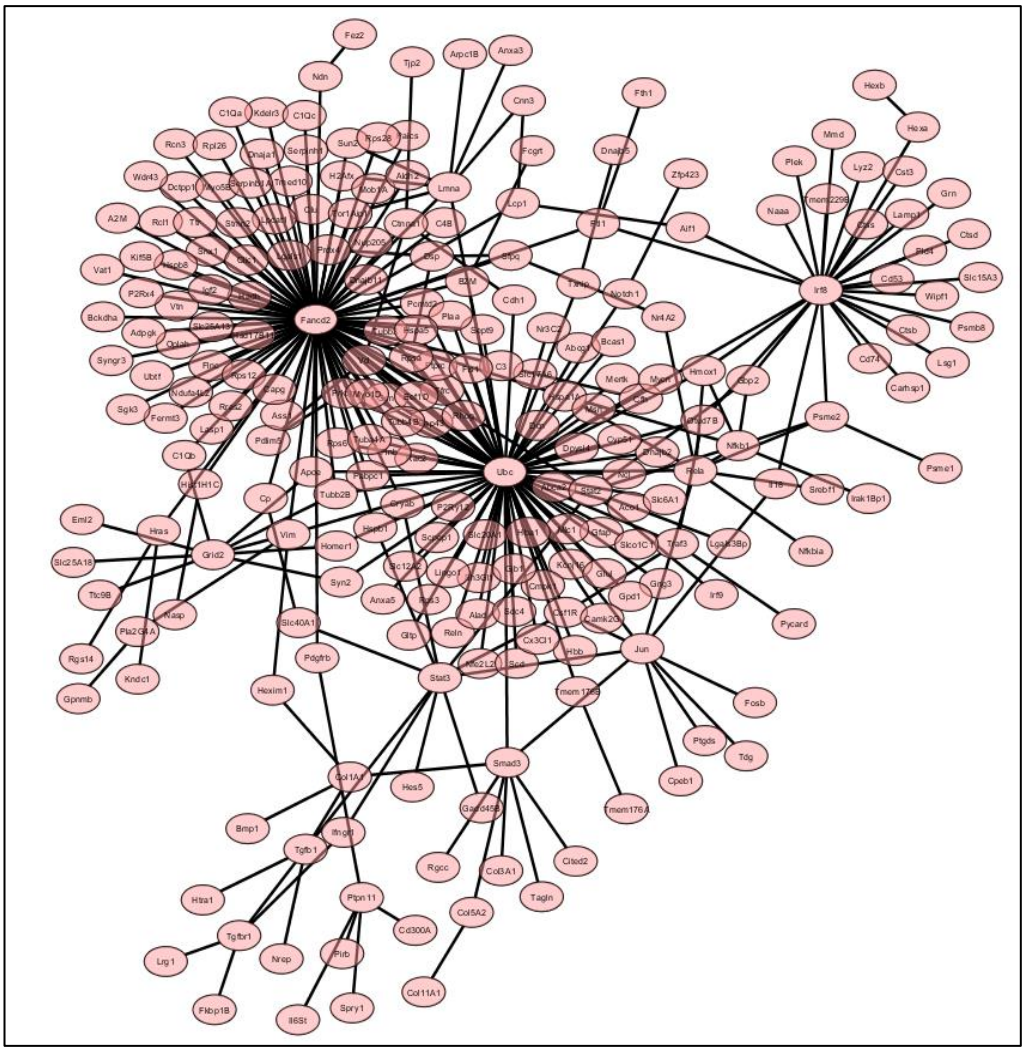

Figure 4.6: Young-Aged comparison subnetwork obtained by KPM for Dataset II at K=2

KPM analysis result are given in Table 4.2. Irf8 is hub gene for K=0, namely it has the highest number of edges in the subnetwork. When K is changed to 1, the size of the network increased, and hub gene changes to Fancd2. Visualization of subnetworks was performed in Cytoscape. Young-Aged comparison subnetwork is shown in Figure 4.6.

4.4.1. Functional analysis of KPM subnetworks for Dataset II

Functional analysis for Dataset II was performed with g:Profiler, as mentioned in Chapter 3.4.1. The subnetwork discovered at K=2 was chosen for functional analysis. Functional analysis was performed with g:Profiler correction at the 0.05 threshold. Only memory, learning, nervous system related functions are given in the

result (Table 4.3). There is not Learning and Memory terms among the identified terms, but there are terms related to brain and aging such as; aging, nervous system development, generation of neurons. Immune system and signaling related terms are also detected in the subnetwork. Interestingly, stress related terms are identified with significant p-values.

Table 4.3: KPM functional analysis results for GSE5666 Young-Aged comparison.

term type	term name	p-value
BP	response to cytokine	3.81E-08
BP	response to stress	2.93E-14
BP	response to oxidative stress	0.000555
BP	immune system process	2E-14
BP	cytokine production	0.00103
BP	cellular response to cytokine stimulus	2.11E-08
BP	Aging	0.000146
BP	nervous system development	2.77E-08
BP	Neurogenesis	6.56E-10
BP	generation of neurons	3.12E-08
BP	neuron differentiation	0.0000223
BP	Gliogenesis	0.000000013
BP	glial cell differentiation	0.00000198
BP	central nervous system development	0.0000103
BP	neuron development	0.0000101
BP	neuron projection development	0.00000147
BP	regulation of cytokine production	0.000588
BP	regulation of immune system process	8.76E-13
BP	positive regulation of immune system process	0.00000184
BP	regulation of immune response	0.00000331
BP	regulation of nervous system development	0.00000358
BP	regulation of neurogenesis	0.000000283
BP	regulation of signaling	0.00000267
BP	positive regulation of signaling	0.00000543
CC	neuron part	0.0000286
CC	neuron projection	0.0000706
rea	Immune System	0.00000244
rea	Innate Immune System	0.000000012

4.5. BioNet Analysis for Dataset II

The ratmusint interactome data was processed with the BioNet package and converted into the GraphNel format with *ftm2graphNel* function in R. p-values of GSE5666 Young- Aged comparison was imported to R and mapped on the interactome data. Nodes were scored by using *scoreNodes* function, and subnetworks were discovered with *runFastHeinz* function.

Only Young-Aged comparison was performed for this dataset and subnetworks were created for different FDR. Subnetwork Discovery was repeated with three different FDR values (0.001,0.01 and 0.05) for this dataset. The subnetworks were created, and their topological properties were shown in Table 4.4. The first column shows Mapped Network and its properties. Nearly 54% transcriptome data was mapped on interactome data. The Mapped Network has 6333 nodes and 24296 edge and Ubc is hub gene for this network.

Table 4.4: BioNet Results for GSE5666 Dataset.

Bionet	FDR= 0.001	FDR =0.01	FDR =0.05
	Size of networks and highest degree of nodes		
Young (10) – Aged (19) p<0.05=1769 p<0.01=924	77 node-90 edge Ubc: 23 Eed: 19	124 node-165edge Ubc: 40 Eed: 30	203 node- 273 edge Ubc: 59 Eed: 45

4.5.1. Functional analysis of BioNet subnetwork for Dataset II

Functional analysis for BioNet subnetwork were performed for FDR= 0.05 since the size of the module was big enough for functional analysis for this FDR cut-off. Nearly same terms obtained from both BioNet and KPM analysis. Stress, Neurogenesis, neuron development and aging are common functions and there is still not Learning and Memory terms. Immune system processes were detected in this moodule with significant p-values.

Table 4.5: BioNet functional analysis result for GSE5666 dataset Young-Aged comparison.

term type	term name	p-value
BP	immune system process	6.51E-12
BP	cytokine production	1.8E-06
BP	Aging	6.09E-05
BP	nervous system development	2.33E-08
BP	Neurogenesis	3.11E-09
BP	Gliogenesis	4.23E-09
BP	generation of neurons	2.5E-07
BP	glial cell differentiation	1.48E-06
BP	astrocyte differentiation	0.00071
BP	neuron differentiation	2.83E-05
BP	neuron development	0.000104
BP	response to cytokine	1.75E-05
BP	cellular response to cytokine stimulus	8.66E-07
BP	response to stress	2.5E-13
BP	response to oxidative stress	0.00483
BP	neuron projection regeneration	0.00188
BP	immune response	2.28E-06
BP	regulation of cytokine production	9.56E-07
BP	positive regulation of cytokine production	0.000308
BP	positive regulation of signaling	0.000128
BP	negative regulation of signaling	0.00056
BP	regulation of nervous system development	0.000255
BP	regulation of neurogenesis	0.000556
BP	regulation of immune system process	5.92E-13
BP	positive regulation of immune system process	4.44E-07
BP	regulation of immune response	0.000654
CC	neuron projection	2.65E-06
rea	Immune System	0.000016
rea	Innate Immune System	5.75E-07

4.6. Corelation analysis for Dataset II

GSE5666 [Rowe et al. 2007] data was used to create Pearson-correlation based co-expression network. Co-expression matrix was created for 11860 unique genes.

4.6.1. Co-expression analysis in Aged and Young groups

In this part of study Aged and Young co-expression networks were created. Young data consists of NT, 5T and 21PT groups. Firstly, highly correlated co-expression maxrix was created for normalized Young NT data in MATLAB with "corr" function. This co-expression matrix includes values from -1 to 1, corresponding to Pearson Correlation values mentioned in Section 3.6.1. 0.95 threshold was used to find highly correlated interactions. Negative and positive correlation values are both used for this study as mentioned in section 3.6.1.

The same analysis was also performed for NT-Aged, 21PT-Young and 21PT-Aged groups. 0.95 was used as a threshold. Mostly connected highly co-expressed network was extracted for each group. The sizes of the networks are shown in Table 4.6. These networks were created to be used in further analysis, the analysis of decreasing co-expression patterns given in Section 4.6.2.

Table 4.6: Co-expression networks for GSE5666 dataset. These co-expression matrixes are created for Aged and Young group separately. Extracted networks represents the most connected network that was extracted from the biggest network.

Group	Extracted aged	Extracted young
	Size of networks and highest degree nodes	
NT	24926 node - 11904 edge Scn2b: 134 Selt: 123	3418 node - 24454 edge Canx: 222 Impact: 220
21PT	4668 node - 13515 edge LOC682635: 45 Dda1: 45	10562 node - 93238 edge Ap2b1: 241 Bzw1: 239

4.6.2. Decreasing Co-expression Pattern Between Age Groups

In this part of study, a correlation network pattern was aimed to be created between Non-Trained and Trained groups to identify gene pairs that show consistently decreasing correlation from Non-Trained to 5-Day-trained to 21-Day Post-trained groups to understand the molecular effects of training on animals. In this dataset there were mainly two groups; Young and Aged. Each group has three sub-groups, as mentioned in Section 4.6.1. Decreasing co-expression pattens were created for Aged and Young groups separately.

Firstly, the correlation values between highly correlated 11904 gene pairs, which were calculated in Chapter 4.6.1 for NT-Aged group, was also calculated in the other two groups, 5-T-Aged and 21PT-Aged. The correlation value can be either positive or negative, so this analysis was separated into two groups; negative high correlation and positive high correlation. The correlation difference between each successive group was defined to be minimum 0.1 to have a pattern. This means that if the correlation value of one gene pair in NT group decreases more than 0.1 in 5T, this correlation shows significant decrease, and obeys the rule. In negative correlation group, there were 747 gene pairs that showed consistent decrease in absolute correlation between the three groups, from NT to 21PT and obeyed the rule. In positive correlation group, there were 2265 interactions that were steadily descending and obeyed the rule. The two groups were combined after threshold filtering, and the combined network had 2177 nodes and 3010 edges. The network was analyzed in Cytoscape and mostly connected part of the network was extracted (Table 4.7).

For Aged 21PT to 5T to NT decreasing pattern, the correlation values between highly correlated 13515 gene pairs, which were calculated in Chapter 4.6.1 for Aged 21PT group, was also calculated in the other two groups, 5-T and NT. Threshold was defined as 0.1. In negative correlation group, there were 860 gene pairs that showed consistent decrease in absolute correlation between the three groups and in positive correlation group, there were 2943 interactions that were steadily descending and obeyed the rule. The two groups were combined after threshold filtering, and the combined network had 3204 nodes and 3801 edges. The network was analyzed in Cytoscape and mostly connected part of the network was extracted (Table 4.7).

The same analysis was performed for Young samples from NT to 5T to 21PT. In negatively highly correlated NT interactions, there were 450, and in positively higly

correlated NT interactions, there were 1220 interactions that were steadily descending and obeyed the 0.1 decreasing rule. The two groups were combined after threshold filtering, and the combined network had 1965 nodes and 1668 edges. The network was analyzed in Cytoscape and mostly connected part of the network was extracted (Table 4.7). From 21PT to 5T to NT decreasing pattern, threshold was defined as 0.15. In negatively highly correlated 21PT interactions there were 9783 interactions, and 13776 interactions were identified in positively highly correlated 21PT interactions. The two groups were combined after threshold filtering, and the combined network had 6648 nodes and 11779 edges. The network was analyzed in Cytoscape, and mostly connected part of networks was extracted (Table 4.7). Visualization of networks was performed in Cytoscape. Decreasing pattern for Young samples from NT to 21 PT was shown in Figure 4.7.

Table 4.7: Co-Expression Network analysis for GSE5666 dataset. Decreasing pattern observed in NT to 21PT and 21PT to NT groups.

Network / Group	Aged	Extracted Aged	Young	Extracted Young
	Size of networks and highest degree of nodes			
From NT to 21PT	2177 node 3010 edge Serp1: 46 Pcnx1: 45	660 node 1995 edge Serp1: 46 Pcnx1: 45	1965 node 1668 edge Gpm6b: 32 Spryd7: 27	458 node 657 edge Gpm6b: 32 Spryd7: 27
From 21PT to NT	3204 node 3801 edge Tcp11: 24 Gas6: 19	1699 node 2753 edge Tcp11: 24 Gas6: 19	6648 node 11779 edge Pou3f1: 34 Mrps30: 32	4966 node 10675 edge Pou3f1: 34 Mrps30: 32

Figure 4.7: Decreasing correlation pattern was created for Young samples from NT to 5T to 21PT.

4.6.3. Functional Analysis of Correlation Networks for Dataset II

Networks of decreasing pattern was created for Aged and Young groups seperately, and these groups were also divided into two subgroups; 21PT to NT and NT to 21PT. Functional analysis was performed for Aged group by using g:Profiler correction with 0.05 threshold and given at Table 4.8 and Table 4.9. Only learning, memory and nervous system related terms are given in the result. The genes that showed decreasing correlation pattern from NT to 5T to 21PT have common functional terms such as nervous system development, neurogenesis, plasticity and brain but they don't have any terms like Learning and Memory. And, circadian rhythm was detected in NT to 21PT module (Table 4.8, Table 4.9). In the case of decreasing correlation from 21PT to 5T to NT there are less terms identified. These terms are mostly about signaling and immune response.

Table 4.8: g:Profiler functional analysis for NT to 21 PT decreasing correlation pattern for Aged group.

term type	term name	p-value
BP	nervous system development	3.16E-08
BP	central nervous system development	0.00184
BP	brain development	0.00367
BP	Neurogenesis	0.000122
BP	generation of neurons	0.000258
BP	neuron differentiation	0.000182
BP	cell morphogenesis involved in neuron differentiation	0.00863
BP	neuron projection development	0.000923
BP	circadian rhythm	0.0107
BP	cell-cell signaling	4.34E-12
BP	trans-synaptic signaling	4.78E-13
BP	chemical synaptic transmission	3.7E-13
BP	signal release from synapse	0.0000857
BP	neurotransmitter secretion	0.0000857
BP	regulation of nervous system development	0.000504
BP	regulation of neurogenesis	0.00231
BP	regulation of neuron differentiation	0.000123
BP	response to stress	0.000836
BP	positive regulation of neuron differentiation	0.00157
BP	cellular response to stress	0.00427
CC	neuron part	3.13E-33
CC	neuron projection	1.61E-21
CC	neuron to neuron synapse	2.3E-11
hp	Brain very small	0.00328
hp	Atrophy/Degeneration affecting the central nervous system	0.00072
hp	Brain atrophy	0.0031
rea	Neuronal System	1.52E-09
rea	Glutamate binding. activation of AMPA receptors and synaptic plasticity	0.00856

Table 4.9: g: Profiler functional analysis for 21PT to NT decreasing correlation pattern for Aged group.

term type	term name	p-value
BP	intracellular signal transduction	0.0307
BP	response to stress	3.05E-10
BP	inflammatory response	0.00685
BP	response to cytokine	0.000194
BP	regulation of signaling	8.2E-06
BP	regulation of signal transduction	2.13E-05
BP	regulation of signaling receptor activity	1.55E-07
BP	cell surface receptor signaling pathway	0.00556
MF	signaling receptor binding	7E-08
MF	cytokine receptor binding	0.00713
Keg	Neuroactive ligand-receptor interaction	0.0165
Tf	Factor: NeuroD; motif: NNSCWGCTGNSY	0.025
Tf	Factor: NeuroD; motif: NNSCWGCTGNSY; match class: 0	0.025

Same functional analysis was performed for the Young group, which has two correlation networks. Decreasing pattern from 21PT to 5T to NT and from NT to 5T to 21PT. The genes in both networks have some common terms such as nervous system, neuron and signaling, but none of them includes Learning or Memory terms. Interestingly, from NT to 21PT Huntington's and Parkinson's Disease terms exist. Brain aging, and memory deficits are closely related to these diseases. There is also synaptic plasticity term in this analysis. From 21PT to NT, circadian rhythm and aging terms exist (Table 4.10, Table 4.11).

Table 4.10: g: Profiler functional analysis for 21PT to NT decreasing correlation pattern for Young group.

term type	term name	p-value
BP	Aging	1.54E-05
BP	nervous system development	1.1E-07
BP	central nervous system development	8.24E-07
BP	response to oxidative stress	8.78E-08
BP	inflammatory response	3.75E-05
BP	circadian rhythm	0.0229
BP	Neurogenesis	3.81E-05
BP	generation of neurons	0.00021
BP	neuron death	3.31E-05
BP	neuron apoptotic process	0.000204
BP	regulation of neurological system process	0.00351
BP	regulation of neurogenesis	0.00392
BP	cytokine production	0.00139
BP	regulation of cytokine production	7.83E-05
BP	positive regulation of cytokine production	0.000237
BP	response to cytokine	5.92E-11
BP	cellular response to cytokine stimulus	4.19E-07
BP	cell-cell signaling	7.39E-13
BP	synaptic signaling	3.68E-10
BP	trans-synaptic signaling	3.68E-10
BP	anterograde trans-synaptic signaling	8.59E-10
BP	chemical synaptic transmission	8.59E-10
BP	regulation of neuron death	0.000377
BP	cell surface receptor signaling pathway involved in cell-cell signaling	0.000747
BP	transmission of nerve impulse	0.00815
CC	Synapse	3.13E-13
CC	neuron to neuron synapse	9.37E-06
CC	neuron part	4.69E-14
CC	neuronal cell body	1.89E-06
CC	neuron projection	1.27E-10
CC	synapse part	1.69E-14
Keg	Neuroactive ligand-receptor interaction	2.03E-11

Table 4.11: g: Profiler functional analysis for NT to 21PT decreasing correlation pattern for Young group.

term type	term name	p-value
BP	nervous system development	0.000341
BP	regulation of neurotransmitter levels	0.00527
BP	regulation of signaling	0.012
BP	cell-cell signaling	0.000121
BP	synaptic signaling	1.65E-05
BP	trans-synaptic signaling	1.65E-05
BP	anterograde trans-synaptic signaling	5.22E-05
BP	chemical synaptic transmission	5.22E-05
BP	signal release from synapse	0.000213
BP	neurotransmitter secretion	0.000213
BP	regulation of synaptic plasticity	0.00117
BP	regulation of neurotransmitter secretion	0.000984
BP	regulation of neuron projection development	0.00362
CC	neuron part	5.49E-10
CC	neuron projection	2.72E-06
CC	neuron to neuron synapse	0.00012
Keg	Huntington disease	0.0278
Keg	Parkinson disease	9.09E-05
Rea	Neuronal System	0.00672
Rea	Neurotransmitter receptors and postsynaptic signal transmission	0.0294

4.7. Weighted Gene Correlation Network Analysis for Dataset II

4.7.1. Detection and Removal of Outliers

GSE5666 [Rowe et al. 2007] dataset is mainly divided into three groups; NT, 5-Day Training and 21—Day Post Training. In order to ensure compatibility with Subnetwork Discovery analysis, only Young (5T) and Aged (5AI/5AU) samples were used for WGCNA analysis. Also, these samples are similar with Dataset I. In Dataset I, Young, Middle aged and Aged samples were given as an input to the WGCNA. For Dataset II similar age samples were used for WGCNA analysis, and these samples were Young (5T) and Aged (5AI/5AU) groups. Young group has 10 samples and Aged group has 19 samples. For Weighted Gene Correlation Network Analysis (WGCNA)

the normalized data was given as an input to the algorithm. WGCNA has its own sample cluster algorithm. The threshold for clustering was chosen as 20 and this threshold led to one sample outlier (5T-AU7). This sample was removed from further analysis (Figure 4.8).

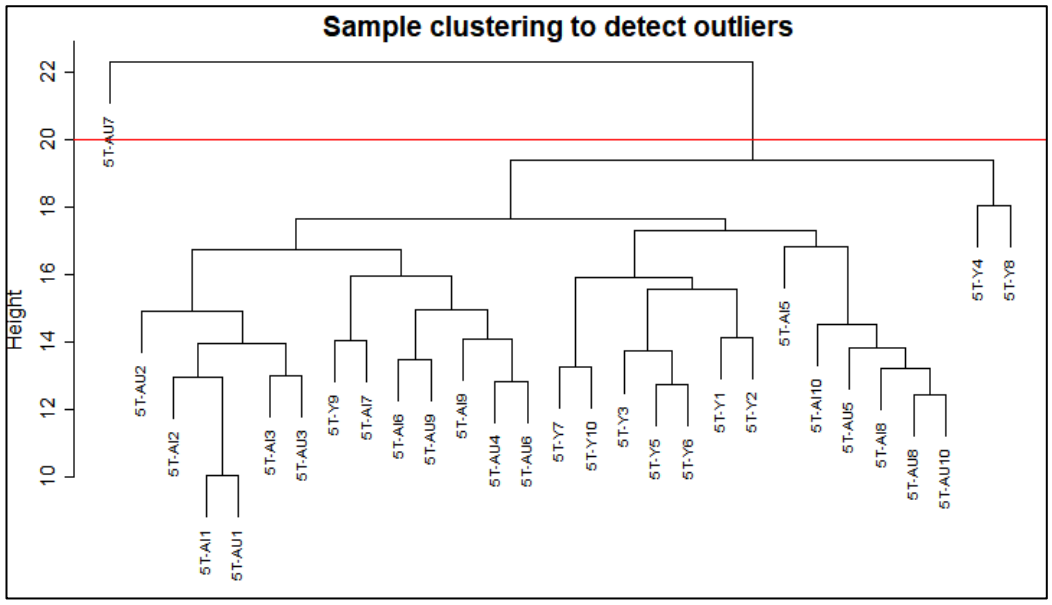

Figure 4.8: Sample clustering tree for GSE854 dataset. The sample 5T-AU7 is an outlier.

4.7.2. Correlating Transcriptome Samples with Phenotypic Trait Data

GSE5666 had information about SWM performance (latency) for each animal (sample). In original study for Dataset II, average latency performance was given for each group. In this thesis study, average latency performance was assigned for each sample. In WGCNA analysis, 5T AI and 5T AU samples were merged to create Aged samples, so the average latency performance was assigned for these samples (5T AI latency = 18.7, 5T AU latency= 8.1, Averaged Aged latency= 13.4). This trait value was introduced to WGCNA package to detect its relationship with transcriptomic data. Moreover, age information was added as a new trait condition to this trait data to associate modules with age (Table 4.12). Each age group (Young and Aged) was defined as a trait data for each sample. “Young” trait is specific to Young samples (10 samples) only, and their trait value was defined as “1”. “Aged” trait is specific to Aged samples (19 samples), and their trait value was defined as “0”. Trait values of samples

can be plotted as a heatmap, and the heatmap can be compared with the dendrogram of transcriptomic samples (Figure 4.9).

Table 4.12: Trait data for GSE5666 dataset. This table consist of latency trait data and the categorical data for “Young” and “Aged” age groups.

Samples	Latency	Young	Aged
5T-Y1	5.3	1	0
5T-Y2	5.3	1	0
5T-Y3	5.3	1	0
5T-Y4	5.3	1	0
5T-Y5	5.3	1	0
5T-Y6	5.3	1	0
5T-Y7	5.3	1	0
5T-Y8	5.3	1	0
5T-Y9	5.3	1	0
5T-Y10	5.3	1	0
5T-AI1	13.4	0	1
5T-AI2	13.4	0	1
5T-AI3	13.4	0	1
5T-AI4	13.4	0	1
5T-AI5	13.4	0	1
5T-AI6	13.4	0	1
5T-AI7	13.4	0	1
5T-AI8	13.4	0	1
5T-AI9	13.4	0	1
5T-AI10	13.4	0	1
5T-AU1	13.4	0	1
5T-AU2	13.4	0	1
5T-AU3	13.4	0	1
5T-AU4	13.4	0	1
5T-AU5	13.4	0	1
5T-AU6	13.4	0	1
5T-AU7	13.4	0	1
5T-AU8	13.4	0	1

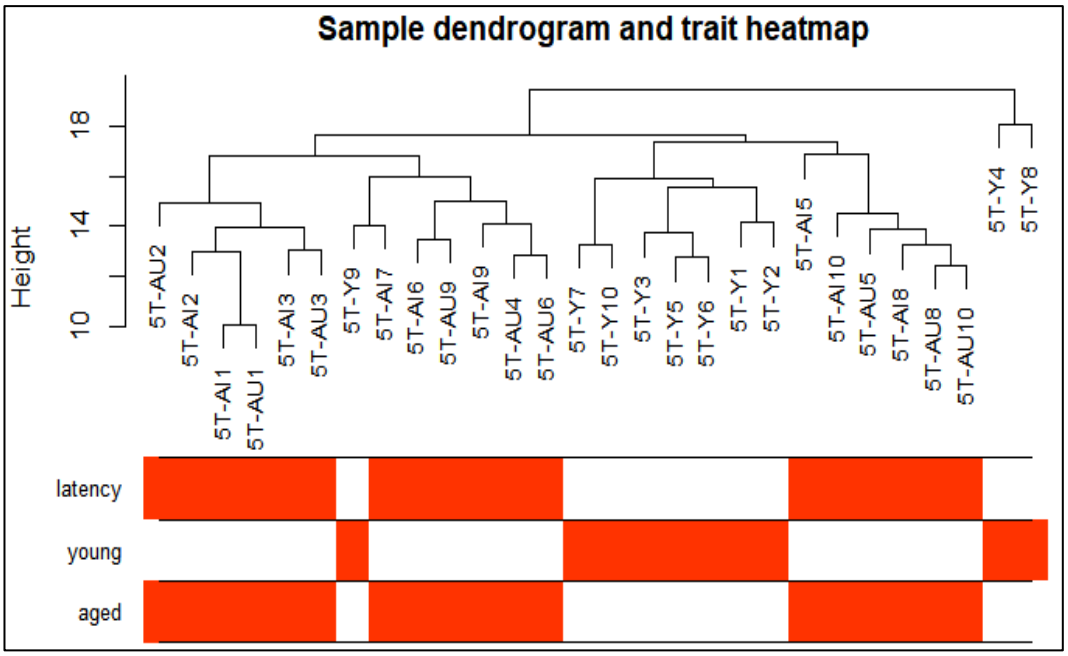

Figure 4.9: Sample dendrogram and trait heatmap plot for GSE5666 dataset. This figure shows the relationship between the transcriptomic samples and the traits.

4.7.3. Soft-Threshold Determination

The next step in WGCNA package is choosing a soft threshold (power) to make co-expression network scale free. With *pickSoftThreshold* function a suitable soft-threshold can be choosen value between 0 and 20. Soft-threshold was chosen as 10, which corresponded to 95% scale free rate. This parameter was used to make the co-expression network scale free by raising the power of each correlation value to the chosen parameter (in this case, cor=cor^{10}) (Figure 4.10).

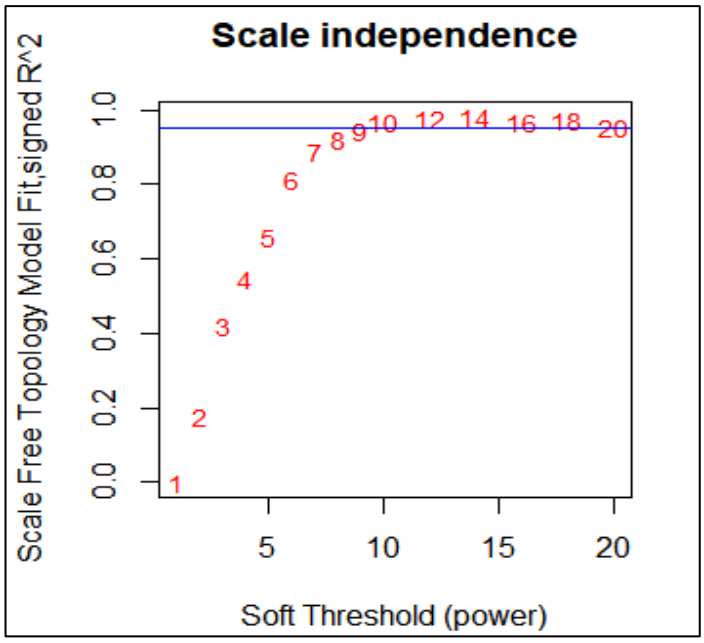

Figure 4.10: Determination of the power parameter for GSE5666 dataset. The model was fitted at 10 to make co-expression matrix scale free in 95% percentage.

4.7.4. Module Creation

The next step is creating modules by using expression data and power. Modules are created by *blockwiseModules* function. This function requires expression data and the power parameter. There are some other parameters important for this analysis; networkType changed to "signed", minModuleSize was adjusted to 40, mergeCutHeight was adjusted to 0.2, other parameters were set to default. Modules were created with these parameters. There were 28 modules created. 1202 genes do not belong to any module, and they were assigned to grey module (Figure 4.11). The module color and their number of nodes were shown in Table 4.13.

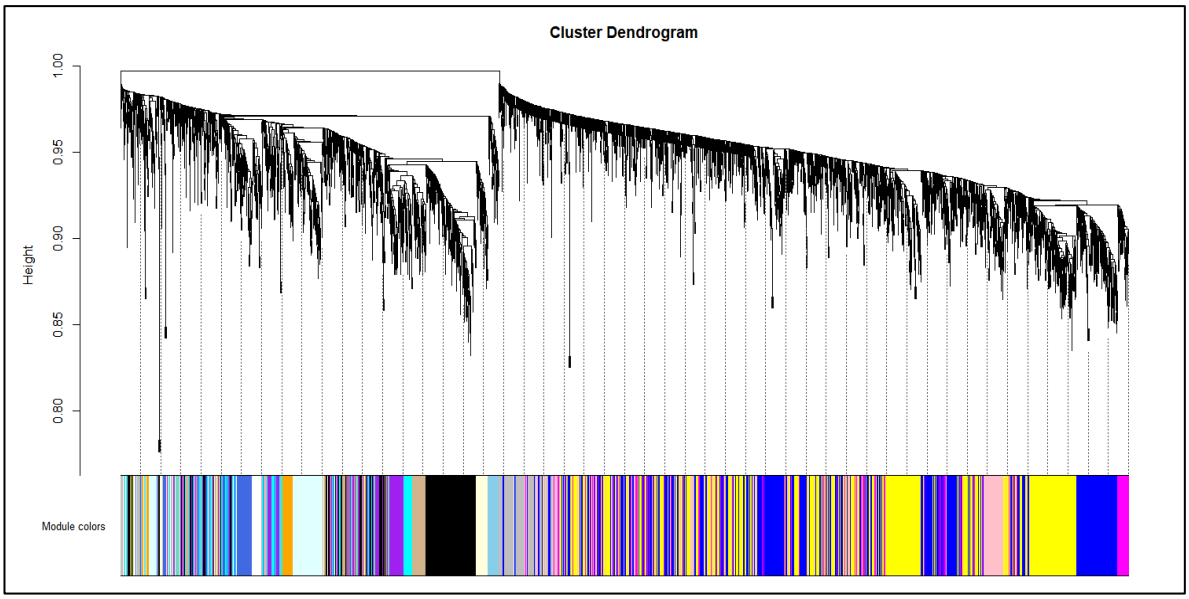

Figure 4.11: Cluster dendrogram for GSE5666 dataset. Modules are created and assigned to a single color. There were 28 modules and each of them had 40 nodes at least.

Table 4.13: Color of modules and their number of nodes.

Color of module	Number of nodes	Color of module	Number of nodes	Color of module	Number of nodes
Black	430	Greenyellow	237	Red	1039
Blue	1142	Grey60	157	Royalblue	116
Brown	1137	Lightcyan	191	Salmon	231
Cyan	209	Lightgreen	148	Skyblue	54
Darkgreen	109	Lightyellow	141	Tan	232
Darkgrey	91	Magenta	308	Turquoise	1238
Darkorange	67	Midnightblue	208	White	64
Darkred	111	Orange	78	Yellow	1125
Darkturquoise	104	Pink	316	Grey	1202
Green	1103	Purple	284		

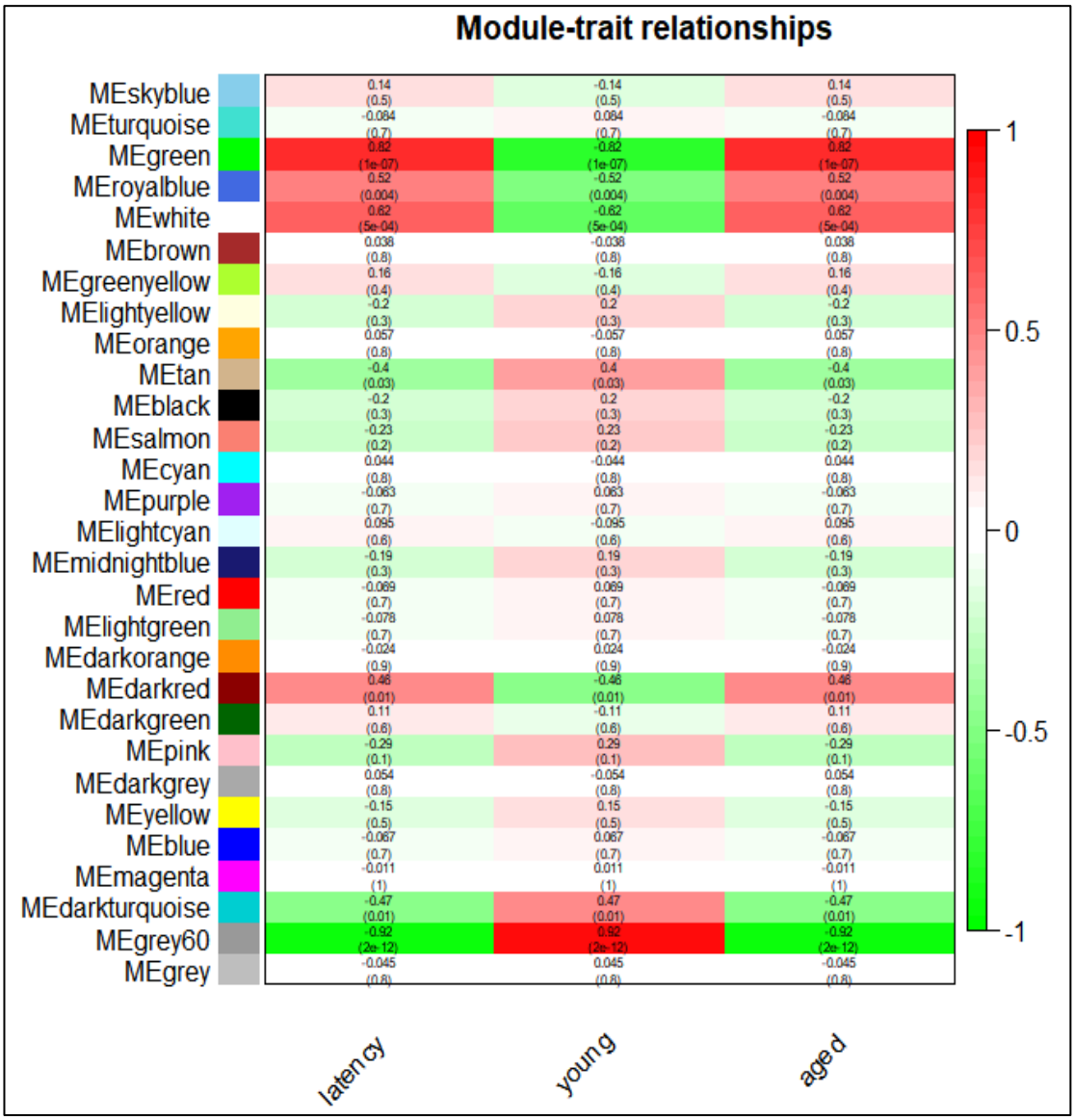

Figure 4.12: Module-trait relationship for Dataset II. This figure shows the relationship of modules to each trait condition.

Module trait relationship graph can be created by using trait data. (Figure 4.12). Royalblue, white and green modules show high correlation with two traits, Aged and Latency traits, and grey60 module is important for Young trait condition. Module trait relationship graph can be used to find important modules in dozens of modules. Module-Membership versus Gene Significance graphs were created for royalblue, white and green modules (Figure 4.13, 4.14 and 4.15). Green module is highly correlated module with two traits. GS & MM graph proved that since Module memberhip shows the correlation of each gene with module and with each other, and GS shows correlation of that module eigengene with traits. Correlation between GS and MM is 0.82, which means significant correlation. The genes in the module are

highly correlated with themselves and they are highly correlated with the latency and aged traits. GS was only plotted for latency trait here, for green, royalblue and white modules.

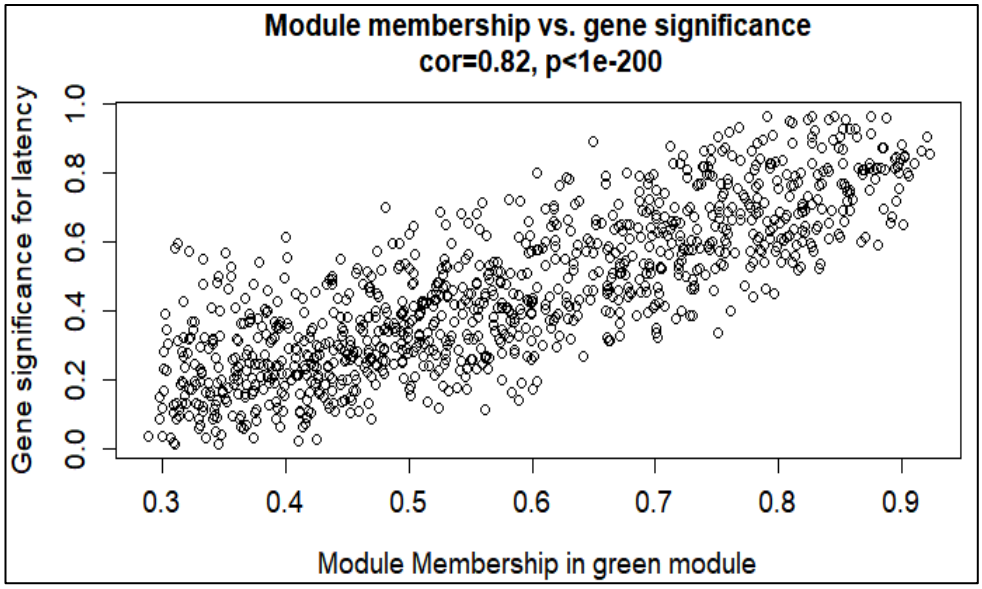

Figure 4.13: GS versus MM graph for green module.

Royalblue and white modules are also significant modules based on latency performance, but the correlation between GS and MM is not significant enough. Membership of nodes in these modules are not correlated with the Gene significance based on the latency performance. Also, genes in these modules are not correlated with each other and with the module significantly. These modules are not significant based on GS versus MM graphs.

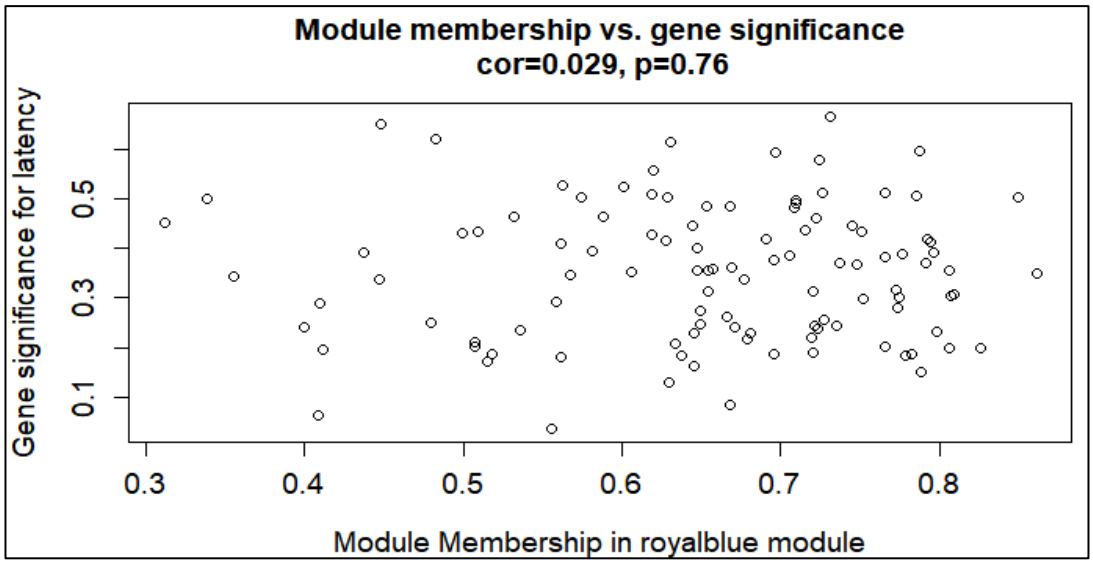

Figure 4.14: GS versus MM graph for royalblue module.

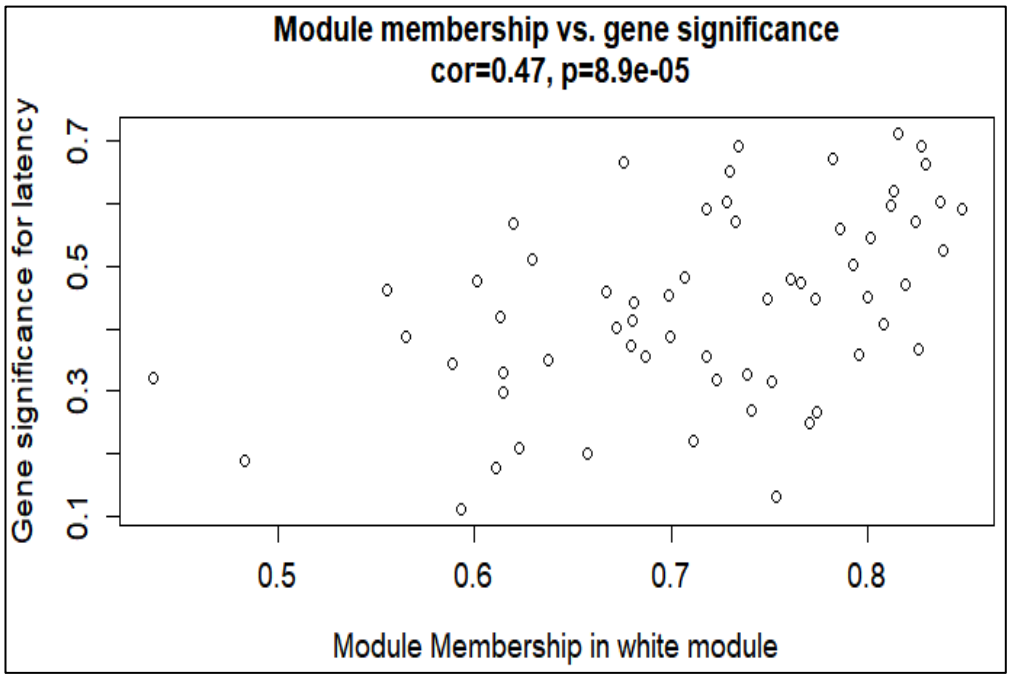

Figure 4.15: GS versus MM graph for white module.

4.7.5. Functional Analysis of Results by using UserListEnrichment function

Functional analysis of the modules was performed using *userListEnrichment* function in WGCNA package (Table 4.14). This function was also used for Dataset I and memory related modules were detected (Table 3.16). Red, grey60, turquoise

modules were identified to be enriched in terms of synaptic plasticity, learning, memory and cognition. These modules are further analysed with g:Profiler.

Table 4.14: *userListEnrichment* function analysis result. This table contains color and number of genes information for important modules based on their significance (Corrected p-value). Corrected p-value gives information about significance of modules for that functional category. These lists were created for this study and they are related to brain, learning and memory.

Module color	User Defined Categories	Corrected p-values
Red	Synaptic plasticity	3.53E-16
Red	Brain	1.99E-11
grey60	Gliogenesis and neurogenesis	8.18E-06
grey60	Brain	5.09E-05
Red	Gliogenesis and neurogenesis	0.010898
Red	Cognition	0.013339
Turquoise	Brain	0.03281
Red	Learning and memory	0.040922

4.7.6. Functional analysis of WGCNA modules for Dataset II

Based on the UserListEnrichment analysis Red, Turquoise and Grey60 modules were analysed with g:Profiler with 0.05 threshold. Only the memory, learning, plasticity related terms are given in the result tables. Red and Turquoise modules include all terms that we are interested in. These modules include Learning, Memory and Cognition and plasticity, and have nervous system and neuron related terms (Table 4.15, Table 4.16)

Table 4.15: g:Profiler results for Red module.

term type	term name	p-value
BP	Behavior	4.94E-06
BP	Cognition	0.000659
BP	learning or memory	0.000743
BP	nervous system development	4.36E-13
BP	Neurogenesis	2.1E-07
BP	generation of neurons	3.71E-07
BP	neuron differentiation	5.67E-07
BP	neuron projection development	1.08E-07
BP	regulation of nervous system development	1.55E-08
BP	negative regulation of nervous system development	0.000102
BP	positive regulation of nervous system development	4.36E-05
BP	cell-cell signaling	4.43E-14
BP	synaptic signaling	9.92E-14
BP	regulation of synaptic plasticity	5.95E-08
BP	regulation of neurogenesis	4.25E-09
BP	regulation of neuron differentiation	1.85E-10
BP	regulation of neuron projection development	3.64E-11
BP	negative regulation of neurogenesis	0.000071
BP	negative regulation of neuron differentiation	3.78E-05
BP	positive regulation of neurogenesis	5.24E-05
CC	neuron to neuron synapse	1.39E-07
CC	neuron part	1.53E-26
CC	neuron projection	2.18E-20
CC	myelin sheath	3.07E-05
Rea	Neuronal System	2.13E-06
Rea	Neurotransmitter receptors and postsynaptic signal transmission	0.0171

Table 4.16: g:Profiler results for Turquoise module.

term type	term name	p-value
BP	cell-cell signaling	1.95E-05
BP	nervous system development	1.92E-12
BP	Neurogenesis	3.05E-07
BP	generation of neurons	2.4E-07
BP	neuron differentiation	0.000215
BP	regulation of nervous system development	0.00104
BP	central nervous system development	0.00297
BP	neuron development	0.00126
BP	neuron projection development	0.000293
BP	regulation of neurogenesis	0.00529
BP	regulation of signaling	0.0168
BP	Behavior	0.0341
BP	Cognition	0.00568
BP	learning or memory	0.00169
BP	Learning	0.0373
BP	associative learning	0.0188
CC	neuron part	3.46E-11
CC	neuron projection	1.24E-09
CC	Axon	1.2E-07
CC	neuronal cell body	0.000212
hp	Abnormality of nervous system physiology	0.0336
keg	Neurotrophin signaling pathway	0.000391
rea	Neuronal System	0.00319

Grey60 module is not good enough in terms of functions but it includes neuron and related terms. But this module includes neurogenesis and neuron development terms and axons, which are parts of neuron, and its production. Green module was found to be important in Module Trait relationship graph. Therefore, it was analysed functionally also. This module does not include Learning and memory terms, but it is the only module that has aging and myelination terms (Table 4.17, Table 4.18).

Table 4.18: g:Profiler results for Green module.

term type	term name	p-value
BP	ensheathment of neurons	0.00227
BP	axon ensheathment	0.00227
BP	Myelination	0.00167
BP	oxidation-reduction process	0.00701
BP	Aging	0.00243
BP	inflammatory response	0.0226
BP	cytokine production	0.0006
BP	regulation of cytokine production	7.44E-05
BP	response to cytokine	0.000103
BP	cellular response to cytokine stimulus	0.00316
BP	regulation of signaling	0.00187
BP	positive regulation of signaling	0.000096
BP	Gliogenesis	1.01E-05
BP	glial cell differentiation	0.000872
BP	glial cell development	2.44E-05
rea	Immune System	7.32E-10

Table 4.17: g:Profiler results for Grey60 module.

term type	term name	p-value
BP	nervous system development	2.83E-06
BP	Neurogenesis	7.41E-09
BP	neuron differentiation	9.85E-08
BP	neuron development	2.74E-07
BP	cell morphogenesis involved in neuron differentiation	0.000521
BP	neuron projection development	1.34E-08
BP	axon development	7.06E-06
BP	neuron projection morphogenesis	1.09E-05
BP	Axonogenesis	9.51E-05
BP	neuron projection guidance	4.19E-05
BP	axon guidance	3.66E-05

5. DISCUSSION

The main objective of this study is to reveal memory and cognition related molecular changes in brain in response to aging by integrating transcriptome data with molecular interaction networks. For this purpose, transcriptome data of two studies [Blalock et al., 2003], [Rowe et al., 2007] were computationally analyzed with two approaches, Subnetwork Discovery and Network Inference. Four algorithms were used to apply these approaches.

5.1. Comparison of Subnetwork Discovery Approaches

Subnetwork analysis was performed by using two methods; BioNet [Beisser et al., 2010] and KeyPathwayMiner [Alcaraz et al., 2014]. These two methods have some advantages and disadvantages over each other. Number of non-significant genes to be included in the subnetwork can be specified with KPM, and visualisation can also be done easily. On the other hand, p-values are accepted by KPM in binarized format. This makes two significant but different p-values the same. For example, p-values of 0.001 and 0.000000001 returns "1" in the binarization. A similar categorization is also valid for non-significant genes. BioNet has only one parameter (FDR) and considers p-values not in binarized format. This algorithm gives each node a weight by using their p-values. But in BioNet the number of allowed non-significant nodes can not be specified by user. The functional analysis shows that desirable modules were created in both algorithms but KPM results have more related functional terms about Memory, Learning and Cogniton. Nearly all KPM modules for Dataset I include more learning and memory terms than BioNet modules. But these terms are less significant in KPM than BioNet. For Dataset I the size of the subnetworks is bigger in KPM than BioNet. For Dataset II, the functional analysis for two algorithms did not give learning and memory terms directly, but BioNet modules have more significant terms than KPM results. KPM gives larger modules than BioNet in young-aged comparison. The number of significantly changed genes in BioNet modules is less than KPM modules, because KPM allows user to define number of non-significat genes.

5.1.1. The effect of threshold on the size of the subnetworks

Bionet and KPM have parameters to define significance of modules. In BioNet, there is one parameter called False Discovery Rate, and this parameter is used as a threshold while scoring nodes. In this study for, BioNet analysis this threshold is changed from 0.01 to 0.2. This value was changed based on the size of the discovered subnetwork; if there is not any identified subnetwork at FDR of 0.01, this value is increased until asubnetwork is detected However, as FDR increases, significance decreases, and false positive values increase. To control the increase of false positive values, an FDR value of more than 0.2 was not used in this study, that is, a maximum of 20 false positive values were accepted in 100 positive values.

Different FDR values were used in BioNet to understand the effect of FDR on the size of the subnetwork. For Dataset I, for young-aged comparison, FDR was increased from 0.01 to 0.2 and the change in the size of identified subnetworks is shown in Table 5.1 and Figure 5.1. In this study 0.1 was chosen as a threshold for young-aged comparison in Dataset I.

Table 5.1: The change of size of the subnetwork based as a function of FDR threshold for Dataset I.

FDR	Number of Nodes	Number of Edges	Hub 1	Degree of hub 1	Hub 2	Degree of hub 2
0.01	0	0	-	-	-	-
0.02	16	16	Ubc	13	-	-
0.03	30	35	Ubc	22	-	-
0.04	51	72	Ubc	31	Eed	14
0.05	68	101	Ubc	42	Slc2A4	26
0.06	90	145	Ubc	51	Slc2A4	35
0.07	103	165	Ubc	58	Slc2A4	38
0.08	125	207	Ubc	69	Eed	33
0.09	152	268	Ubc	79	Eed	37
0.1	**219**	**419**	**Ubc**	**91**	**Slc2A4**	**64**
0.15	410	846	Ubc	143	Slc2A4	103
0.2	645	1403	Ubc	215	Slc2A4	151

In Table 5.1 with the increase in FDR the size of the subnetwork increases. The first hub first mostly connected gene, which has the highest number of interactions, of the subnetworks did not change, and only the number of connected nodes (degree) of the hub gene increased.

BioNet considers the degree of genes and their p-values while scoring nodes to discover subnetworks, so the hub gene is the same in all identified subnetworks. Only second mostly connected node and their degree changes based on the FDR. The distribution of subnetworks is shown in Figure 5.1.

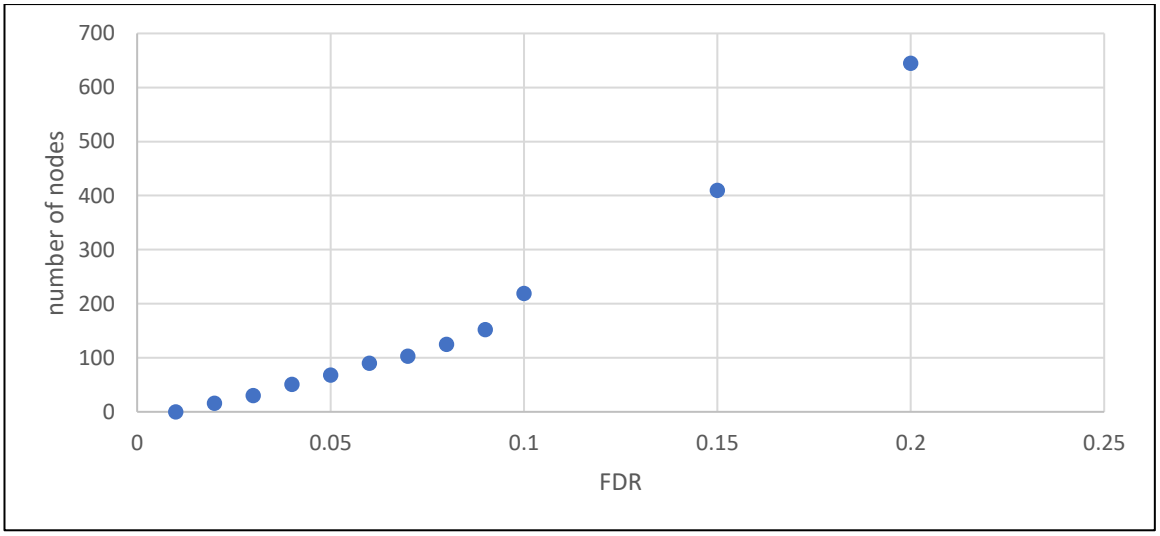

Figure 5.1: Network size change based on FDR for BioNet analysis of Dataset I.

KeyPathwayMiner accepts significance of change in a binarized format as input, where 0 shows no change, and 1 shows significant change for each gene. Based on the defined threshold (p-value of 0.01 or 0.05) p-values of genes are chaned into 0 or 1 and given as input to the KPM algorithm. In KPM there are two parameters that can be changed by user. The first one is p-value threshold. In this study, 0.01 or 0.05 are used as a threshold. The effect of change of this threshold is shown in Table 5.2.

Table 5.2: The change of size of the subnetwork based on p-value threshold and K value for Dataset I.

p-value	K: 0 Network size	K: 0 hubs and degree	K: 2 Network size	K: 2 hubs and degree	K: 4 Network size	K: 4 hubs and degree
0.001	24n-24e	App: 6 Grin1: 6	72n-101e	Ubc: 45 Fancd2: 32	87n-151e	Ubc: 46 Fancd2: 32
0.005	24n-24e	App: 6 Grin1: 6	72n-101e	Ubc: 45 Fancd2: 32	87n-151e	Ubc: 46 Fancd2: 32
0.01	44n-50e	Grin1: 9	106n-156e	Ubc: 62 Slc2a4: 43	124n-206e	Ubc: 63 Fancd2: 43
0.02	102n-139e	Hnrnpk: 22 Ywhaz: 20	183n-299e	Ubc: 92 Slc2a4: 63	218n-391e	Ubc: 94 Fancd2: 69
0.03	172n-253e	Gja1: 36 Hnrnpk: 30	255n-462e	Ubc: 111 Fancd2: 91	293n-567e	Ubc: 113 Fancd2: 91
0.04	239n-367e	Gja1: 43 Hnrnpk: 34	329n-613e	Ubc: 128 Fancd2: 113	365n-787e	Ubc: 129 Fancd2: 114
0.05	271n-421e	Gja1: 49 Hnrnpk: 37	375n-700e	Ubc: 142 Fancd2: 128	415n-892e	Ubc: 143 Fancd2: 129

The p-value was changed from 0.001 to 0.05 at constant K value (K=0). The size of the subnetwork increased with the changing p-value because the significance is decreased and the probability that a significant gene is in the module is increasing. The distribution of networks based on changing p-value is plotted for K=0 and shown in Figure 5.2.

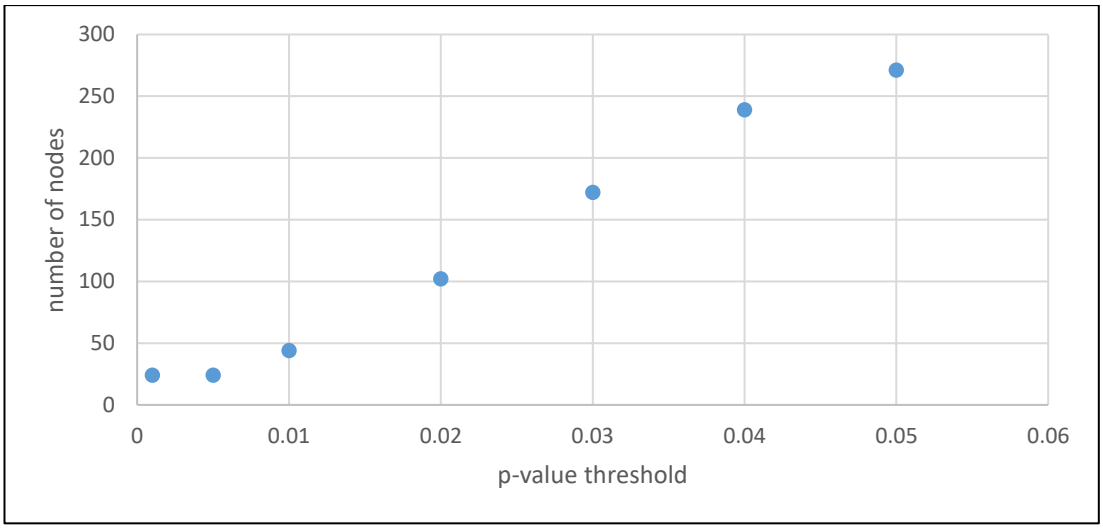

Figure 5.2: Size change of network based on changing p-value threshold for KPM analysis at K=0 (Dataset I).

The second important parameter for KPM is K value, which shows number of non-significant genes in the discovered subnetwork. The effect of changing K on the size of subnetwork at constant p-value is investigated for Dataset I. With the increase in the number of non-significant genes, size of the network increased, and hub genes changed (Table 2). At K=0, hub gene is App and Grin, but when K is changed to 1 KPM choose one non-significant gene that has the most connections - hub gene (Ubc)- and its connections increase the size of the subnetwork. This size change is shown in Figure 5.3.

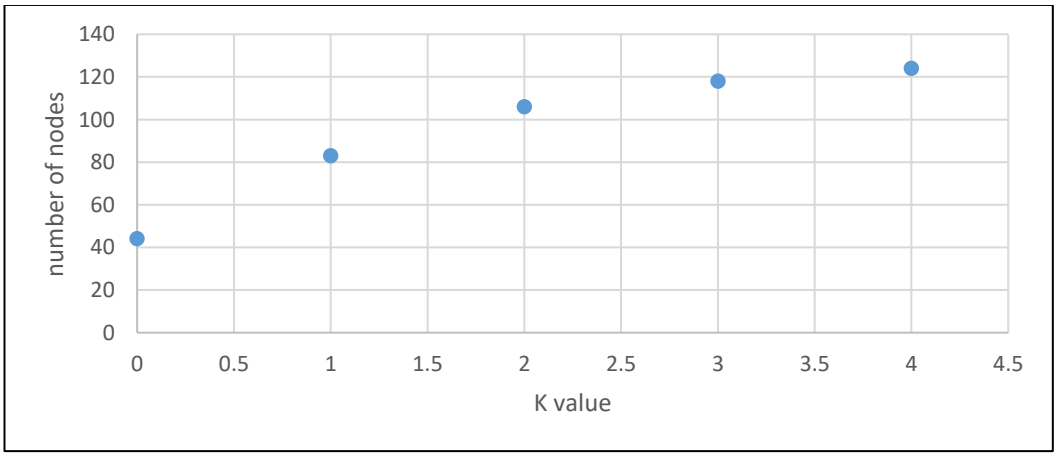

Figure 5.3: Size change of subnetworks based on changing K value for KPM analysis at p=0.01 (Dataset I)

5.1.2. Analysis of BioNet modules

KPM and BioNet are subnetwork discovery algorithms and they detect subnetworks by using transcriptome data and organism specific PPI network. Subnetworks include significant and non-significant nodes. Based on the defined threshold, in KPM the number of non-significant can be defined by "K" value but in BioNet number non-significant genes in the subnetworks can not be defined by user. To compare networks that are detected by two algorithms, the number of non-significant genes available in BioNet subnetworks were checked for Dataset I. The size of the subnetworks and their number of non-significant genes are shown Table 5.3.

Table 5.3: Comparison of the subnetworks for Dataset I

Comparison groups	BioNet module size /FDR	Number of non-significant genes	KPM module size / K	Number of non-significant genes	Common nodes
Dataset I (GSE854)					
Young-Middle aged	113 node-178 edge (fdr=0.1)	2	231 node-365 edge (K=0)	0	78
Middle aged- Aged	379 node-685 edge (fdr=0.2)	42	196 node-264 edge (K=0)	0	196
Young-Aged	219 node-419 edge (fdr=0.1)	89*	44 node-50 edge (K=0)	0	44
Dataset II (GSE5666)					
Young-Aged	203 node- 273 edge (fdr=0.05)	9	240 node-311 edge (K=2)	2	145
*the number of non-significant genes were found based on 0.01 p-value threshold. At p-value threshold of 0.05, the number of non-significant genes is 8.					

BioNet and KPM give different-size subnetworks based on the Table 5.3. For Dataset I there is not any non-significant genes in the KPM modules, and these modules includes memory and learning related genes (Chapter 3). KPM modules are statistically more significant than BioNet modules. BioNet and KPM modules have common genes -nodes - for three different comparisons. In Middle aged - Aged and Young-Aged comparisons, BioNet modules includes all the nodes in KPM modules, and they have additional nodes. BioNet gives more comprehensive but statistically less significant subnetworks than KPM and those modules include all KPM nodes. For Dataset II the size of the subnetworks is more similar than Dataset I.

5.2. Comparison of Network Inference Approaches

Network Inference analyse were performed by using Pearson Correlation and WGCNA algorithm. WGCNA has lots of parameters and optimizing these parameters for our data was time consuming. In this algorithm, modules are created based on their correlation similarity, and each module is represented by a color. This algorithm also accepts trait data to identify modules associated with trait data. In our analysis, on the other hand, functional analysis results show that the modules which were apredicted to be important based on module-trait relationship graphs are not good enough for our analyis, because functional analysis of these modules does not lead to terms about Learning, Memory, Plasticity and other terms such as neurogenesis, nervous system development or aging. Therefore, for this analysis, custom gene lists were created for Memory, Learning, Neurogenesis and Gliogenesis, and userListEnrichment function of WGCNA package was used to detect modules enriched with the genes in those gene lists. Modules were enriched based on the created gene lists. Pearson correlation was used to create an interaction network with decreasing correlation pattern between age groups. These networks were later functionally analysed. These modules have terms related to the learning and memory, and this means that this analysis works well for this case.

5.2.1. Analysis of WGCNA Modules

WGCNA is a Network Inference algorithm and used to identify highly correlated genes and to place these genes in a module. Each Module is represented in a single color. These modules include highly correlated genes, and they are assumed to be in common function. Since the list of significantly changed genes is also used as an input to UserListEnrichment function of WGCNA package, the function detects the number of significant genes in all modules. The significant genes are detected in each module, but, first, significant gene list is created. Significant gene list was created by considering all significantly changed genes in three comparison groups. In Dataset I, there were three comparison groups; Young-Middle aged, Middle aged-Aged and Young- Aged. There were two options to choose p-value threshold, in this part of analysis two of them were used separately, gene lists were created for both. In Young-Middle aged comparison there were 855 significantly changed gene for p=0.05, and 276 genes significantly changed at p=0.01. Other number of significantly changed genes are shown in Table 3.2. These genes are combined and all significant genes in each comparison is used for UserListEnrichment analysis. For Dataset I a total of 2004 genes are significantly changed at p=0.05, 753 genes are significantly changed at p=0.01. Both lists were used, and modules that were significantly enriched with significantly changed genes were detected. Enrichment results are shown in Table 5.4.

Table 5.4: UserListEnrichment analysis of WGCNA modules for Dataset I.

Color of module	User Defined Categories	Corrected p-values	Number of genes
Lightyellow	p-value=0.01	1.70E-29	51
Yellow	p-value=0.01	2.70E-09	68
Lightcyan	p-value=0.01	5.10E-04	31
Lightyellow	p-value=0.05	5.10E-22	67
Yellow	p-value=0.05	1.30E-17	153
Lightcyan	p-value=0.05	5.70E-07	66
Darkgrey	p-value=0.05	0.001377	29
Tan	p-value=0.05	2.54E-02	84

Based on the results in Section 3.7.2 the most important modules were Black, Salmon and Lightcyan. Within these modules Lightcyan is the only one module that either includes memory and learning related genes or includes high number of significantly changed genes. The combination of all analysis results should be considered while choosing the most suitable module in WGCNA analysis. The most suitable module (i) should have higher correlation with trait data, (ii) the genes in that module should be related to memory and learning functions based on the functional analysis and (iii) higher percentage of genes in this module should be changed significantly based on a pre-defined threshold. For Dataset I, Lightcyan is an important module for WGCNA analysis. The same analysis was also performed for Dataset II. There was only young-aged comparison, significantly changed gene list is created at p value=0.01, there were 924 genes significantly changed.

Table 5.5: UserListEnrichment analysis of WGCNA modules for Dataset II.

Color of module	User Defined Categories	Corrected p-values	Number of genes
Green	p-value=0.01	4,08E-294	491
grey60	p-value=0.01	2,10E-76	104
White	p-value=0.01	5,86E-10	24
royalblue	p-value=0.01	0,005760834	21

Based on UserListEnrichment result for Dataset II, Green, Grey60, White and Royal blue modules have high number of significantly changed genes. Grey60 module was found to be important for WGCNA analysis in Section 4.7.3.

5.3. Functional Analysis of Subnetworks and Modules

Functional analysis of the subnetworks and modules, that were created by four different algoritms, includes terms such as Learning, Memory, Plasticity, Cognition and other nervous system related functions. These terms provide an indirect validation of algorithms since they are related to the learning and memory mechanisms, based on the literature. There were some other related terms also detected in the results. Cytokine is one of them and known to be important in the immune system. It was

shown that cytokines and their related mechanisms have learning and memory-related functions [Donzis et al., 2014]. Cytokines are generally known to be associated with immune system and in neuronal processes. Any manipulation of these processes is related to inflammation and affects neuronal systems. Cytokines and their downstream signaling cascades have an impact on the modulation of learning and memory [Donzis et al., 2014].

Gliogenesis is another term that were identified in the functional analysis. Neurogenesis have functions in neuronal connectivity in specific brain areas, but gliogenesis generates oligodentrocytes and astrocytes for myelination. They both help brain to produce neurons and glia in mature brain. Altered gliogenesis and neurogenesis mechanisms may affect the central nervous system and cause neuropsychiatric and neurodegenerative diseases [Rusznák et al., 2016]. Myelination is also important for learning and memory; any disorder causes a decline in learning mechanisms. Myelin is a sheet that covers the axons and protects them from the environment. With myelin sheet, the electrical transmission gets faster and energy is saved. Myelin is affected by the emergence of redundant myelin profiles with normal aging. Any problem in myelination causes a problem in signal transmission [Nave et al., 2014].

A circadian rhythm is a 24-hour cycle in the physiological processes of living beings. Circadian rhythm is controlled by internal stimulus, but it can also be modulated externally with sunlight and temperature. With this internal timing system, organisms can react to external effects. Timing system also affects the learning mechanisms. For example, mice maze memory performances increase in the dark phase of the light. Hippocampus is important in learning and memory, and circadian rhythm is also regulated by hippocampus. It is important for many processes in the body, and any disorder causes neurodegenerative diseases such as Alzheimer's, Parkinson's, and Huntington's diseases [Smarr et al., 2014].

Cell-cell signaling was also detected in the functional annotation results. Intercellular signaling is actively present in many processes in body. Matter and energy is transmitted from one cell to another to carry information with the help of signaling processes within the cell. Since the brain is controlled by the exchange of information between neurons, any disorders in intercellular signaling has an important function in the presence of neurological disorders. Any disorders in this signaling mechanism causes lack of information and communication between cells [Koseska et al., 2017].

This study reveals that these two approaches and four algorithms are appropriate for transcriptome data. Subnetwork Discovery approaches can detect right subnetworks from the huge PPI data and Network Inference approaches create meaningful correlated modules.

This analysis can also create a new point of view to the molecular mechanisms of learning and memory processes. The subnetworks have related terms about learning and memory, but they also have other terms that are not yet known to be related to memory. Among those are Lipid mechanism, Calcium balance, intracellular and extracellular transport etc. If these terms are analyzed experimentally, the effect of aging on learning and memory can be revealed more comprehensively.

5.4. Comparison of Datasets

In this study, the effect of aging on learning and memory performance was aimed to be elucidated. There are several studies in literature that used rats from different age groups, trained them with SWM and other cognitive and memory tests, and Hippocampus region of their brain was extracted, and basic statistical analyses were performed. GSE854 [Blalock et al. 2003] and GSE5666 [Rowe et al., 2007] are two datasets used in this study. The two datasets use *Rattus norvegicus* as a model organism. In both datasets, animals are trained with MWM test, and ages of samples are close to each other. There are three trained age groups in Dataset I (GSE854), which are Young (4 months old), Middle aged (14 months old) and Aged (24 months old), and there are two trained age groups in Dataset II, which are Young (4-6 month old) and Aged (26 months old), and Dataset II has also non-trained Young (4-6 months old) and Aged (24-26 months old) animals. Two datasets have common samples for Young and Aged groups. These samples are comparable for Subnetwork Discovery and WGCNA analysis.

These studies used same organisms and same region of the brain (hippocampal CA1 region), but different platform was used for transcriptome analysis. GSE 5666 used Affymetrix Rat Expression 230A Array and GSE854 used Affymetrix Rat Genome U34 Array. The number of genes that bind to these chips are different for the same organism and the same region of the brain, therefore, the number of the gene expression is different for two datasets. In Dataset I 5749, and in Dataset 11860 unique genes and their expression data exist. Dataset II includes nearly twice as much data

than Dataset I. The number of significant changed genes based on the comparison but for Young-Aged comparison is also different. Dataset I has 407 (7% of data) significantly changed genes at p=0.01 threshold and Dataset II has 924 (8% of data) significantly changed genes at the same threshold. Nearly same percentage of data significantly changed for two datasets.

KPM and BioNet algorithms map transcriptome data on interactome data. Mapped genes create Mapped Network and this network was used for discovering case specific significant subnetworks. For Dataset I in KPM, 61% of the transcriptome is mapped on interactome, and 37% of the interactome is used. Nearly 3500 genes were used from transcriptome data for Subnetwork Discovery analysis. For Dataset II 72% of the transcriptome data and 58% of the interactome was used. Nearly 8500 genes were used for subnetwork discovery analysis. The size of the subnetworks is higher in Dataset II based on Dataset I (Table 3.2-Table 4.2) because the number of mapped genes is higher.

Functional analysis was performed for the identifieed subnetworks in g:Profiler, with g:Profiler correction at 0.05 threshold. In Subnetwork analysis, Dataset I gives more related terms about learning and memory than Dataset II. They both have neuron, nervous system related terms, cytokine production and regulation, immune system processes, neurogenesis and gliogenesis functional terms, but in Dataset I for young aged comparison synaptic plasticity, learning and memory terms exist in functional analysis.

In Network inference analysis two datasets were used in different ways. In Dataset I the effect of aging is investigated but in Dataset II training effect is investigated for aged and young group separately. In WGCNA analysis the same type of data was used, in both computational analyses the effect of aging on memory performance is investigated. WGCNA modules are functionally analyzed with UserListEnrichment function and based on the results Dataset II modules have more related terms about learning and memory mechanism than Dataset I. They both have neurogenesis, neuron part and signaling terms, but Dataset I has memory, learning and cognition terms additionally and modules in Dataset II has higher number of significantly changed genes (Table 5.4-5.5).

As a result, Dataset I gives more reliable subnetworks in Subnetwork Discovery analysis based on functional analysis, but Dataset II gives more related and significant terms in functional analysis of WGCNA modules.

5.5. Novelty of Network-based Data Analysis Over the Traditional Analysis

Several experimental designs in literature are created to understand the effect of aging on learning and memory performance. Generally, *Rattus norvegicus* is used as a model organism in these studies, and animals with different age groups are trained with SWM, OMT or other memory tests. At the end of the training, hippocampal region of brain of animals is extracted, and transcriptome data is collected. Basic statistical analyses (Student t-test, ANOVA) are performed with transcriptome data and differentially expressed genes are identified and functional analysis is performed to detect the underlying mechanisms of the effect of aging on memory.

In this study, these analyses were expanded with network-based approaches. Data were obtained from learning and memory experiments and a few statistical tests were performed. Mainly two network-based approaches were used for this study, Subnetwork discovery, which maps transcriptome data on organism specific PPI network to find subnetworks, and Network Inference, which creates modules by using Pearson correlation. Network based approaches are important to understand topological relationships of genes and their functions in specific conditions. Not only significant genes but also experimentally proved interaction-relationships of genes were used to understand the effect of aging on memory deficits.

Subnetwork Discovery and Network Inference give subnetworks and modules. Functional analyses of these modules proved that these modules include cognition, learning and memory related terms. Some of these terms were already identified in the original studies that reported those transcriptome data, without incorporating networks into analysis (Rowe et al., 2007 and Blalock et al., 2003). These terms are nervous system development, immune system processes, signal transduction, axonal growth, myelinogenesis, cytokine production and regulation. In the network-based analysis, learning, memory, synaptic plasticity, circadian rhythm terms are clearly detected, which are important functions for learning and memory. All these terms are related to learning and memory based on the literature (Chapter 5.3). There are other terms detected with the functional analysis of subnetworks or modules, which are response to organic substance, response to external stimulus, apoptotic process, protein binding, cell migration and motility. These terms need further investigation to associate them directly with learning and memory mechanisms.

5.6. Critical Assessment of Functional Analysis Results

In this study, two network-based approaches were applied to two transcriptomic datasets, and modules and subnetworks were created. The quality of the subnetworks based on their functional analysis shows some differences. Some modules include significantly changed genes and some of them have high correlation with trait data, but they have not related terms about learning and memory based on their functional analysis. To understand the reason of these findings better, genes in grey60 modules were analysed one by one. This module was created by using WGCNA algoritm by using Dataset II (GSE5666). It was found to be correlated with Young trait data (Figure 4.12) and included gliogenesis and neurogenesis terms based on UserListEnrichment analysis (Table 4.14). This module also has higher number of significantly changed genes based on the UserListEnrichment result (Table 5.5). But, g:Profiler analysis of this module did not give satisfying results in terms of common functions of the included genes (Table 4.17). Therefore, the genes in the modules were analyzed separately to understand if their functions are related to memory and learning mechanism.

Grey 60 module has 157 genes. In g:Profiler results, learning and memory related mechanisms, which include neuron development, gliogenesis and neurogenesis mechanisms and other signaling mechanisms, were detected. Genes that belong to these mechanisms were excluded from further analysis to focus on the remaining genes to check if they have memory and learning associated functions reported in literature bu not included in GO and Pathway enrichment databases. 39 genes were detected that have function in memory and learning related mechanisms. Remaining genes were searched in NCBI and GeneCards to understand their function. Camkk2 (Calcium/Calmodulin Dependent Protein Kinase) gene is one of the memory and learning related genes. This gene is strongly expressed in brain and effects signalling processes about learning and memory mechanisms, and has function in synapse formation, which is important for synaptic plasticity [Peters et al., 2003].

Chst10 (Carbohydrate Sulfotransferase 10) gene was also detected in the grey60 module, and it is not associated with memory and learning related function in g:Profile analysis result. This gene is responsible for the production of HNK-1 protein which has role in synaptic plasticity in hippocampus. This protein also has function in neurodevelopment [Senn et al., 2002]. Znrf1 (Zinc and Ring Finger 1) gene is also

detected in Grey60 module. This gene plays role in neuron cell differentiation and maintenance of neuronal transmission, and it is related to plasticity [Araki et al., 2003].

This analysis shows that absence of the information on the association of some genes with learning and memory related mechanisms in the enrichment databases can be the reason behind why some modules are not related to memory and learning mechanisms based on the functional analysis. g:Profiler or other functional analysis tools may not include up-to-date information for each gene and their function, leading to some poor predictions.

REFERENCES

Albert R., (2007), "Network Inference, Analysis, and Modeling in Systems Biology", The Plant Cell, 19 (1), 3327-3338.

Alcaraz N., Kucuk H., Weile J., Wipat A., Baumbach J., (2011), "KeyPathwayMiner: Detecting Case-Specific Biological Pathways Using Expression Data", Internet Mathematics, 7 (4), 299-313.

Alcaraz N., Pauling J., Batra R., Barbosa E., Junge A., Christensen A., Azevedo V., Ditzel H., Baumbach J., (2014), "KeyPathwayMiner 4.0: condition-specific pathway analysis by combining multiple omics studies and networks with Cytoscape", BMC Systems Biology, 8 (99), 1-19.

Antoniadis E., Ko C., Ralp M., McDonald R., (2000), "Circadian rhythms, aging and memory", Elsevier, 111 (1-2), 25-37.

Araki T., Milbrandt J., (2003), "ZNRF proteins constitute a family of presynaptic E3 ubiquitin ligases", Journal of Neuroscience, 23 (28), 9385-9394.

Aranda B., Blankenburg H., Kerrien S., Brinkman F., Ceol A., Chautard E., Dana J., De Las Rivas J., Dumousseau M., Galeota E., Gaulton A., Goll J., Hancock R., Isserlin R., Jimenez R., Kerssemakers J., Khadake J., Lynn D., Michaut M., O'Kelly G., Ono K., Orchard S., Prieto C., Razick S., Rigina O., Salwinski L., Simonovic M., Velankar S., Winter A., Wu G., Bader G., Cesareni G., Donaldson I., Eisenberg D., Kleywegt G., Overington J., Ricard-Blum S., Tyers M., Albrecht M., Hermjakob H.,(2011), "PSICQUIC and PSISCORE: accessing and scoring molecular interactions", Nat. Methods, 29 (8), 528-9.

Batra R., Alcaraz N., Gitzhofer K., Pauling J., Ditzel H., Hellmuth M., Baumbach J., List M., (2017), "On the performance of de novo pathway enrichment", npj Systems Biology and Applications, 3 (6), 1-19.

Baudry M., Bi X., (2009), "Synaptic Plasticity: Learning and Memory in Normal Aging", Encyclopedia of Neuroscience, 757-762.

Beisser D., Klau G., Dandekar T., Müller T., Dittrich M., (2010), "BioNet: an R-Package for the functional analysis of biological networks", Bioinformatics, 26 (8), 1129-1130.

Blalock E., Chen K., Sharrow K., Herman J., Porter N., Foster T., Landfield P., (2003), "Gene Microarrays in Hippocampal Aging: Statistical Profiling Identifies Novel Processes Correlated with Cognitive Impairment", The Journal of Neuroscience, 23 (9) 3807-3819.

Buechel H., Popovic J., Staggs K., Anderson K., Thibault O., Blalock E., (2014), "Aged rats are hypo-responsive to acute restraint: implications for psychosocial stress in aging, Frontiers", 6 (13), 1-16.

Barbulovic-Nad I., Lucente M., Sun Y., Zhang M., Wheeler A., Bussmann M., (2006), "Bio-Microarray Fabrication Techniques—A Review", Critical Reviews in Biotechnology, 26 (4), 237-259.

Chen J., Zhang N., Wen J., Zhang Z., (2017), "Silencing TAK1 alters gene expression signatures in bladder cancer cells", Oncology Letters, 13 (5), 2975-2981.

Cheng L., Li L., Whang L., Li X., Xing H., Zhou J., (2018), "A random forest classifier predicts recurrence risk in patients with ovarian cancer", Molecular Medicine Reports, 18 (3), 3289-3297.

Clough E., Barret T., (2016), "The Gene Expression Omnibus database", Methods Mol. Biol., 1418, 93-110.

Crowder R., (2015), "Principles of Learning and Memory", Classic Edition, Psychology Press.

Dittrich MT., Klau G., Rosenwald A., Dandekar T., Müller T., (2008), "Identifying functional modules in protein–protein interaction networks: an integrated exact approach", Bioinformatics, 24 (13), 223–231.

Donzis E., Tronson N., (2014), "Modulation of learning and memory by cytokines: signaling mechanisms and long-term consequences", Neurobiol Learn Mem., 115, 68-77.

Dutta R., Chomyk A., Chang A., Ribaudo M., Deckard S., Doud M., Edberg D., Bai B., Li M., Baranzini S., Fox R., Staugaitis S., Macklin W., Trapp B., (2013), "Hippocampal demyelination and memory dysfunction are associated with increased levels of the neuronal microRNA miR-124 and reduced AMPA receptors", Ann Neurol., 73 (5), 637-45.

Gene Ontology Consortium, (2019), "The Gene Ontology Resource: 20 years and still GOing strong", Nucleic Acids Research, 47 (D1), D330–D338.

Gersner J., Yin J., (2010), "Circadian rhythms and memory formation", Nature, 11, 577–588.

Giurgiu M., Reinhard J., Brauner B., Dunger-Kaltenbach I., Fobo G., Frishman G., Montrone C., Ruep A., (2019), "CORUM: the comprehensive resource of mammalian protein complexes", Nucleic Acids Research, 47 (D1), 559-563.

Heinbockel T., (2017)," Synaptic Plasticity", Thomas Heinbockel Edition.

Hemminger B., (2005), "Introduction to the Special Issue on Bioinformatics", Journal of the American Society for Information Science and Technology, Third Edition.

Horvath S., Zhang Y., Langfelder P., Kahn R., Boks M., Eijk K., Berg H., Ophoff H., (2012), "Aging effects on DNA methylation modules in human brain and blood tissue", 13 (10), R97.

Huang X., Lin X., Zeng J., Wang L., Yin P., Zhou L., Hu C., Yao W., (2017), "A Computational Method of Defining Potential Biomarkers based on Diferential Sub-Networks", Scientific Reports, 7 (14339), 1-10.

Johansson B., (2012), "Multisensory stimulation in stroke rehabilitation", Frontiers in Human Neuroscience, 6, 60.

Kalamohan K., Gunasekaran P., Ibrahim S., (2019), "Gene coexpression network analysis of multiple cancers discovers the varying stem cell features between gastric and breast cancer", Elsevier, 21 (100576), 1-24.

Kaptan Z., Üzüm G., (2016), "The Role of Adult Hippocampal Neurogenesis in Learning and Memory Function", Turk J Neurol, 22, 149-155.

Koseska A., Bastiaens P., (2017), "Cell signaling as a cognitive process", The EMBO Journal, 36, 568-582.

Krumsiek J., Suhre K., Illig T., Adamski J., Theis F., (2011) "Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data", BMC Systems Biology, 5 (21), 1-16.

Kusonmano K, (2016), "Gene Expression Analysis Through Network Biology: Bioinformatics Approaches", Adv Biochem Eng Biotechnol, 160, 15-32.

Langfelder P., Horvath S., (2007), "Eigengene networks for studying the relationships between co-expression modules", BMC Systems Biology, 1 (54), 1-17.

Langfelder P., Horvath S., (2008) "WGCNA: an R package for weighted correlation network analysis", BMC Bioinformatics, 9 (559), 1-13.

Lazarus S., (1991), "Cognition and motivation in emotion", American Psychologist, 46 (4), 352-367.

Lesk A., (2019), "Introduction to Bioinformatic", Fifth Edition, Oxford University Press

Licata L., Briganti L., Peluso D., Perfetto L., Iannuccelli M., Galeota E., Sacco F., Palma A., Nardozza A., Santonico E., Castagnoli L., Cesareni G., (2011), "MINT, the molecular interaction database: 2012 update", 40 (D1), D857–D861.

Llinás M., Bozdech Z., Wong E., Adai T., DeRisi J., (2006), "Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains", Nucleic Acid Research, 34 (4), 1166–1173.

Love S., (2006), "Demyelinating diseases", Journal of Clinical Pathology, 59 (11), 1151-1159.

Malone J., Oliver B., (2011), "Microarrays, deep sequencing and the true measure of the transcriptome", BMC Biology, 9 (34), 1-9.

Miller J., Olham M., Geschwind D., (2008), “ASystems Level Analysis of Transcriptional Changes in Alzheimer’s Disease and Normal Aging”, The Journal of Neuroscience, 28 (6), 1410-1420.

Mitra T., Menon S., Sinha S., (2018), “Emergent memory in cell signaling: Persistent adaptive dynamics in cascades can arise from the diversity of relaxation timescales”, Scientific Reports, 8 (13230), 1-15.

Mitre M., Mariga A., Chao M., (2017), “Neurotrophin signalling: novel insights into mechanisms and pathophysiology”, Clin Sci (Lond), 131 (1), 13-23.

Brito-Moreira J., Lourenco MV., Oliveira MM., Ribeiro FC1., Ledo JH., Diniz LP., Vital JFS., Magdesian MH., Melo HM., Barros-Aragão F., de Souza JM., Alves-Leon SV., Gomes FCA., Clarke JR., Figueiredo CP., De Felice FG., Ferreira ST., (2017), “Interaction of amyloid-β (Aβ) oligomers with neurexin 2α and neuroligin 1 mediates synapse damage and memory loss in mice”, J Biol Chem., 292 (18), 7327-7337.

Nave K., Werner H., (2014), “Myelination of the Nervous System: Mechanisms and Functions”, Annu. Rev. Cell Dev. Biol., 30, 503-533.

Orchard S., Ammari M., Aranda B., Breuza L., Briganti L., Broackes-Carter F., Campbell N., Chavali G., Chen C., Toro N., Duesbury M., Dumousseau M., Galeota E., Hinz U., Iannuccelli M., Jagannathan S., Jimenez R., Khadake J., Lagreid A., Licata L., Lovering R., Meldal B., Melidoni A., Milagros M., Peluso D., Perfetto L., Porras P., Raghunath A., Ricard-Blum S., Roechert B., Stutz A., Tognolli M., Roey K., Cesareni G., Hermjakob H., (2014), “The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases”, Nucleic Acids Research, 42 (D1), D358–D363.

Perlman R., (2016), “Mouse models of human disease”, Evolution, Medicine, and Public Health, (1), 170–176.

Rankin C., Beck C., Chiba C., (1990), “Caenorhabditis elegans: A new model system for the study of learning and memory”, ELSEVIER, 37 (1), 89-92.

Peters M., Mizuno K., Ris L., Angelo M., Godaux E., Giese K., (2003), “Loss of Ca2+/Calmodulin Kinase Kinase β Affects the Formation of Some, But Not All, Types of Hippocampus-Dependent Long-Term Memory”, Journal of Neuroscience, 23 (30), 9752-9760.

Rasch B., Born J., (2013), “About Sleep’s role in Memory”, Physiological Rewievs, 93 (2), 681-766.

Razick S., Magklaras G., Donaldson I., (2008), “iRefIndex: A consolidated protein interaction database with provenance”, MC Bioinformatics, 9 (405), 1-19.

Reimand J., Arak T., Adler P., Kolberg L., Reisberg S., Peterson H., Vilo J., (2016), “g:Profiler—a web server for functional interpretation of gene lists”, Nucleic Acids Research, 44 (W1), 83-89.

Rusznák Z., Henskens W., Schofield E., Kim W., Fu Y., (2016), "Adult Neurogenesis and Gliogenesis: Possible Mechanisms for Neurorestoration", Experimental Neurobiology, 25 (3), 103-112.

Rowe W., Blalock E., Chu Chen K., Kadish I., Wang D., Barrett J., Thibault O., Porter N., Rose G., Landfield P., (2007), "Hippocampal Expression Analyses Reveal Selective Association of Immediate-Early,Neuroenergetic,and Myelinogenic Pathways with Cognitive Impairment in Aged Rats", The Journal of Neuroscience, (12), 3098-3110.

Salzer J., Zalc B., (2016), "Myelination", Science Direct, 26 (20), R971-R975.

Senn C., Kutsche M., Saghatelyan A., Bösl MR., Löhler J., Bartsch U., Morellini F., Schachner M., (2002), "Mice deficient for the HNK-1 sulfotransferase show alterations in synaptic efficacy and spatial learning and memory", Elsevier, 20 (4), 712-729.

Shannon P., Markiel A., Ozier O., Baliga N., Wang J., Ramage D., Amin N., Schwikowski B., Ideker T., (2003), "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks", Genome Research, 13 (11), 2498-2504.

Smarr B., Jennings K., Driscoll J., Kriegsfeld L., (2014), "A Time to Remember: The Role of Circadian Clocks in Learning and Memory", 128 (3), 283-303.

Stark C., Breitkreutz B., Reguly T., Boucher L., Breitkreutz A., Tyers M.,(2006), "BioGRID: a general repository for interaction datasets", Nucleic Acids Research, 34 (1), D535–D539.

Suri D., Bhattacharya A., Vaidy V., (2013), "Early stress evokes temporally distinct consequences on the hippocampal transcriptome, anxiety and cognitive behavior", International Journal of Neuropsychopharmacology, 17 (2), 289-301.

The UniProt Consortium, (2019), "UniProt: a worldwide hub of protein knowledge", Nucleic Acids Research, 47 (D1), D506–D515.

Verbitsky M., Yonan A., Malleret G., Kandel E., Gilliam C., Pavlidi P., (2004), "Altered Hippocampal Transcript Profile Accompanies an Age-Related Spatial Memory Deficit in Mice", Learning & Memory, 11 (3), 253-256.

Web 1, (2018), https://keypathwayminer.compbio.sdu.dk/keypathwayminer/, (Date of access: 20/02/2018).

Web 2, (2018), https://bioconductor.org/biocLite.R , (Date of access: 24/08/2018).

Web 3, (2019), http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Rpackages/WGCNA, (Date of access: 18/03/2019).

Web 4, (2018), https://www.ensembl.org/biomart/martview, (Date of access: 16/03/2018).

Web 5, (2018), https://peterlangfelder.com/2018/11/25/working-with-categorical-variables/, (Date of access: 25/11/2018).

Xue J., Schmidh S., Sander J., Draffenh A., Krebs W., Quester I., Nardo D., Gohel T., Emde M., Schmidleithner L., Ganesan H., Nino-Castro A., Mallmann M., Labzin L., Theis H., Kraut M., Beyer M., Latz E., K Schultze J., (2014), "Transcriptome-Based Network Analysis Reveals a Spectrum Model of Human Macrophage Activation", Science Direct, 40 (2), 274-288.

Yin J., Liu X., Wang B., Wang D., Wei M., Fang H., Xhang M., (2016), "Gene expression profiling analysis of ovarian cancer", Oncology Letters, 12 (1), 405-412.

Yirmiya R, Goslen I., (2011), "Immune modulation of learning, memory, neural plasticity and neurogenesis", Elsevier, 25 (2), 181-213.

Warsow G., Struckmann S., Kerkhoff C., Reimer T., Engel N., Fuellen G., (2013), "Differential Network Analysis Applied to Preoperative Breast Cancer Chemotherapy Response", Plos One, 8 (12), e81784.

Werhli A., Grzegorczyk M., Husmeier D., (2006), "Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks", Systems biology, 30 (9), 1236-40.

Zerbino D., Achuthan P., Akanni W., Amode1 R., Barrell D., Bhai1 J., Billis K., Cummins C., Gall A., Gir´on C., Gil L., Gordon L., Haggerty L., Haskell E., Hourlier T., Izuogu1 O., Janacek S., Juettemann T., To1 J., Laird M., Lavidas I., Liu Z., Loveland J., Maurel1 T., McLaren W., Moore B., Mudge J., Murphy D., Newman V., Nuhn M., Ogeh D., Ong C., Parker A., Patricio M., Riat H., Schuilenburg H., Sheppard D., Sparrow H., Taylor K., Thormann A., Vullo A., Walts B., Zadissa A., Frankish A., Hunt S., Kostadima M., Langridge N., Martin1 F., Muffato M., Perry E., Ruffier M., Staines D., Trevanion S., Aken B., Cunningham F., Yates A., Flicek P., (2018), "Ensembl 2018", Nucleic Acids Research, 46 (D1), D754-D761.

Zhang B., Horvath S., (2005), "A general framework for weighted gene co-expression network analysis", Stat Appl Genet Mol Biol., 4 (1), 1-43.

Zhang J., An J., (2007), "Cytokines, Inflammation and Pain", Int Anesthesiol Clin., 45 (2), 27-37.

BIOGRAPHY

Elif EMANETCİ was born in Sarıyer in September 24, 1992. She graduated from Marmara University Department of Bioengineering in 2015. She is an M.Sc. student in Bioinformatics and Systems Biology Program under Bioengineering Department at Graduate School of Natural and Applied Sciences in the same university.