TRANSCRIPTOMIC NETWORK ANALYSIS OF BRAIN AGING AND ALZHEIMER'S DISEASE

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

POORYA PARVIZI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
BIOLOGY

AUGUST 2017

Approval of the thesis:

TRANSCRIPTOMIC NETWORK ANALYSIS OF BRAIN AGING AND ALZHEIMER'S DISEASE

submitted by **POORYA PARVIZI** in partial fulfillment of the requirements for the degree of **Master of Science in Biology Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Orhan Adalı _____
Head of Department, **Biology**

Assoc. Prof. Dr. Mehmet Somel _____
Supervisor, **Biology Dept., METU**

Assist. Prof. Dr. Nurcan Tunçbağ _____
Co-supervisor, **Health Informatics, METU**

Examining Committee Members:

Prof. Dr. Tolga Can
Computer Engineering Dept., METU _____

Assoc. Prof. Dr. Mehmet Somel
Biology Dept., METU _____

Assist. Prof. Dr. Nurcan Tunçbağ
Health Informatics, METU _____

Assoc. Prof. Dr. Michelle M. Adams
Psychology Dept., İhsan Doğramacı Bilkent University _____

Assoc. Prof. Dr. A. Elif Erson Bensan
Biology Dept., METU _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: POORYA PARVIZI

Signature :

ABSTRACT

TRANSCRIPTOMIC NETWORK ANALYSIS OF BRAIN AGING AND ALZHEIMERS DISEASE

Parvizi, Poorya
M.S., Department of Biology
Supervisor : Assoc. Prof. Dr. Mehmet Somel
Co-Supervisor : Assist. Prof. Dr. Nurcan Tunçbağ

August 2017, 74 pages

Multiple studies have investigated aging brain transcriptomes to identify for age-dependent expression changes and determine genes that may participate in age-related dysfunction. However, aging is a highly complex and heterogeneous process where multiple genes contribute at different levels depending on individuals' environments and genotypes. Both this biological heterogeneity of aging, as well as technical biases and weaknesses inherent to transcriptome measurements, limit the information gained from a single data set. Here we propose using network analysis to reproducibly identify aging-related gene interactions shared across different datasets. We employ the prize-collecting Steiner forest algorithm to create aging networks on human brain transcriptome datasets. The algorithm calculates the optimal interaction set among aging-related genes within a protein-protein interaction (PPI) network, taking into consideration expression-age correlation coefficients of the most differentially expressed genes with age, and the PPI confidence scores. This allows aging-related genes to interact either directly or through intermediate nodes. The interme-

diate nodes, in turn, can represent genes undetected in transcriptome data due to low light intensity, technical inefficiency of platforms, or aging-related molecular changes that do not involve mRNA abundance change, such as aging-related post-translational modifications. Using the predicted networks, we have performed network alignment of the reconstructed networks to test whether common interactions might be found in different tissues' aging networks. In addition, we also extend the approach to compare molecular changes during aging and in Alzheimer's Disease. We hypothesize that using network alignment will help highlight the most relevant gene clusters and pathways shared between the two processes.

Keywords: Aging, Alzheimer's diseases, Transcriptome, Aging Network, Network alignment, Prize-collecting Steiner forest

ÖZ

BEYİN YAŞLANMASI VE ALZHEİMER HASTALIĞI'NIN TRANSKRİPTOMİK AĞ ANALİZİ

Parvizi, Poorya
Yüksek Lisans, Biyoloji Bölümü
Tez Yöneticisi : Doç. Dr. Mehmet Somel
Ortak Tez Yöneticisi : Yrd. Doç. Dr. Nurcan Tunçbağ

Ağustos 2017, 74 sayfa

Birçok çalışma, yaşa bağlı ekspresyon değişimlerini belirlemek ve yaşa bağlı fonksiyonel bozukluklara katılan olası genleri tespit etmek için beyin yaşlanma ifadesi çalışmıştır. Ancak, yaşlanma bireylerin çevresine ve genotipine bağlı olarak, birden fazla genin farklı seviyelerde katkıda bulunduğu oldukça karmaşık ve heterojen bir süreçtir. Transkriptom ölçümlerine özgü teknik eğilim ve zayıflıkların yanı sıra yaşlanmanın biyolojik heterojenliği, tek bir veri setinden elde edilen bilgiyi sınırlar. Burada, farklı veri setlerinde paylaşılan, yaşa bağlı gen etkileşimlerinin tekrarlanabilir olarak belirlenmesi için ağ analizi kullanılması gerektiğini ileri sürüyoruz. İnsan beyni transkriptom veri setlerinde yaşlanma ağları oluşturmak için prize-collecting Steiner forest algoritması kullanıyoruz. Algoritma, yaşla birlikte farklı olarak anlatılan genlerin, gen anlatımı-yaş korelasyon katsayılarını ve PPI güven skorlarını göz önüne alarak, bir protein-protein etkileşimi ağı içinde, yaşla ilişkili genler arasındaki optimum etkileşimi hesaplar. Yaşlanma ile ilişkili genlerin doğrudan veya ara nodlar

aracılığıyla etkileşime girmesine izin verir. Ara nodlar, düşük ışık yoğunluğu, platformların teknik olarak efektif olmaması veya yaşla ilişkili translasyon sonrası modifikasyonlar gibi mRNA yoğunluğunu değişimini içermeyen yaşlanmayla ilişkili moleküler değişiklikler nedeniyle transkriptom verilerinde de değişim göstermeyen genleri temsil edebilir. Tahmini ağları kullanarak, farklı dokuların yaşlanma ağlarında ortak etkileşimlerin bulunup bulunmadığını test etmek için yeniden yapılandırılmış ağların ağ uyumluluğunu gerçekleştirdik. Buna ek olarak, yaşlanmada ve Alzheimer Hastalığı'nda moleküler değişiklikleri karşılaştırıyoruz. Ağ hizalaması kullanımının, iki süreç arasında paylaşılan en alakalı gen kümelerine ve yolaklara dikkat çekmeye yardımcı olacağını ileri sürmekteyiz.

Anahtar Kelimeler: Yaşlanma, Alzheimer Hastalığı, Transkriptom, yaşlanma ağı, Ağ hizalaması, ödül-toplama steiner algoritması

To my brother

ACKNOWLEDGMENTS

First and foremost, i would like to express my sincere gratitude to my advisor Mehmet Somel for his motivation, support and share of his immense knowledge since my undergraduate degree. I am also grateful for the friendly environemt he establish in the group, which made easier to learn and discuss scientific subjects.

I would like to express my gratitude to my co-supervisor Nurcan Tunçbağ for her guidance, encouragement and precious feedbacks during this study. I appreciate the opportunity she gave me to be part of her group.

I would like to thank my lab mates from both "comparative evolutionary biology" and "Network modeling" labs for their fruitful suggestions and comments in my research studies and their friendships since my first day in the lab. I must thank to Hamit İzgi, Handan Melike Dönertaş, Ekin Sağlıcan, Dilek Koptekin, Hazal Moğultay and Güngör Budak for sharing their experinces. I must especially thank to Zeliha Gözde Turan for her helps and answering my questions regardless of the time during of my MSc study.

I also want to thank Arif Badem, Arvin Hosseinnejad, Ashkan Soltani, Danial Gorgani, Deniz Sezer, Furugh Raouf, Golden Nadimi, Nazila Farhangzad, Nima Sohrabnia, Orkun Evran and Safoora kamjan for their friendship and encouragement.

Last but not least, i would like to indicate my gratitude to my parents for their patience and endless supports. They have always believed in me, even when I didn't believe in myself. Without the help of them, this study would not be finished. I also grateful to my brother, Payam Parvizi, for his positive attitude, support and undoubted companionship throughout my life.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

1.1 What is Aging

Human-beings have a strong desire for living longer since ancient times. To illustrate, the ancient Romans believed that following healthy diet and living according to it could prolong healthy life and increase life expectancy (Cokayne, 2003). Aging is the dysfunction or changes in biological pathways with respect to time. It is statistically stated that, in the last decade the number of aging population have risen sharply due to the developments in medicine and industry. For this reason, it is believed that the proportion of people over 65 will increase from 15% in 2009 to 26% in 2039 (Hsieh, 2015). Accumulation of cellular damage during aging and increase in the proportion of elderly individuals would raise the prevalence of age-related diseases including cancer, neurodegenerative and cardiovascular diseases (Brunet and Berger, 2014). Therefore, understanding the molecular mechanisms of aging is a valuable approach in the exploration of disease processes.

Caenorhabditis elegans is the premier model organism for aging studies which first introduced by Sydney Brenner in 1963 (Tissenbaum, 2015). He believed that the model organism should be cheap, easily reproducible in the lab, with a short generation time and a simple body plan. Utilizing this biological model, Klass searched for mutant strains of *C. elegans* which could extend its lifespan (Klass, 1983). He found a significant correlation between increase in a life span and food intake. His pioneer research opened a new era in the biology of aging studies. Since then lots of aging studies have applied which most of them categorized in one of hallmarks of aging.

1.2 Hallmarks of Aging

Various changes in molecular pathways and mechanisms can contribute to aging processes. These alterations, together, explain the phenotype of aging. These characteristics, in molecular and cellular level are categorized into nine hallmarks of aging (López-Otín et al., 2013). All of the hallmarks carry three criteria: (1) it should participate in normal aging, (2) its trigger should accelerate aging, (3) its amelioration should increase lifespan.

1.2.1 Genomic Instability

Somatic cells are constantly under exogenous and endogenous threats. These detrimental agents induce DNA lesion in genomic and mitochondrial DNA and defects in nuclear architecture. Amelioration of these DNA damages is surveillanced by DNA repair mechanisms. However, the failure to repair or incorrect repair could lead to instability and increase in mutation rate in a cell (Vijg and Suh, 2013). Mutation accumulation is one of the main factors in aging (Moskalev et al., 2013). Although, DNA damage accumulation is the cause of premature aging disease, like progeria syndrome, the association of this disease and aging is unclear.

1.2.2 Telomere Attrition

Telomere is a repetitive DNA sequence found at the end of chromosomes to protect it against attrition and fusion with other chromosomes. Telomerase, the enzyme that adds repetitive nucleotides to the 3' end of telomere is not express in human somatic cells and some other mammalian cells. Time dependent telomere exhaustion cease the cell proliferation and leads to the cellular senescence (Hayflick and Moorhead, 1961; Olovnikov, 1996). In addition, experiments of mice exhibit the decrease in lifespan in telomere shortened samples. On the other hand, induction of telomerase activity extend longevity in mice (Armanios et al., 2009; Blasco et al., 1997; Herrera et al., 1999; Rudolph et al., 1999; Tomás-Loba et al., 2008).

1.2.3 Epigenetic Alterations

Epigenetic changes in aging, projected in transcriptomic alterations and disruption of genome architecture (Brunet and Berger, 2014). Changing in DNA methylation, histone modifications and chromatin remodeling result in genome instability, malfunction in DNA repair mechanism and increase in transcriptional noise (Pal and Tyler, 2016). Various experiments promise the effect of epigenetic alterations in aging processes and onset of premature aging disease. To illustrate, deficient of SIRT6 protein deacetylase, accelerate aging in mice (Mostoslavsky et al., 2006). On the other hand, increased activity of this protein increase the life span (Kanfi et al., 2012).

1.2.4 Loss of Proteostasis

Protein function and their structure are kept under tight surveillance of protein quality control mechanisms to eliminate or ameliorate nonfunctional and incorrectly folded proteins. Unfolded or misfolded proteins, mostly refolded with the help of heat-shock proteins, i.e. chaperones. However, some of them undergo degradation through ubiquitin pathway or engulfed and broken down by lysosomes. Failure to ameliorate problematic proteins results in their accumulation in a cell. Loss of proteostasis during aging is demonstrated by various studies (Koga et al., 2011).

1.2.5 Deregulated Nutrient-sensing

Nutrient-sensing pathways detect nutrient intake and regulate anabolic signaling in a cell according to it. Insulin/IGF-1 signaling (IIS) pathway is contributed to aging process and evolutionarily conserved. Mutation in this pathway and downstream components increase lifespan. It is experimentally proven that caloric restriction which deregulated nutrient-sensing and drugs which mimics nutrient availability increase the healthy aging (Fontana et al., 2010).

1.2.6 Mitochondrial Dysfunction

Decrease in ATP production and disturbance in mitochondrial respiratory chain is one of the features of cellular aging. Reactive oxygen species are byproducts of mitochondrial respirations. It is believed that, accumulation of these free radicals lead to the functional disruption in mitochondria. In addition, these changes could be due to mutation accumulation in mtDNA (Park and Larsson, 2011).

1.2.7 Cellular Senescence

Cellular senescence is an exhaustion of cell proliferation. In addition to the telomere shortening stated by Hayflick, other age-related mechanisms might contribute in this process. Accumulation of these cells, diminish the efficient function of tissue (Campisi and d'Adda di Fagagna, 2007). However, cellular senescence is also helpful in the elimination of cells with abnormal growths and hence protect from tumor formations. Studies claim that, over-activation of tumor suppressor pathways which are induced due to senescence, extend life span (Matheu et al., 2007, 2009). On the other hand, elimination of senescent cells in premature aging model organism delays age-related pathologies (Baker et al., 2011).

1.2.8 Stem cell Exhaustion

Decrease in the potential of stem cells in regeneration through aging is one of characteristics that participate in aging phenotype. Extreme proliferation of the stem cells leads to the stem cell exhaustion resulting in deficiency in regeneration of new cells. In addition, excessive proliferation of intestinal stem cells in Drosophila resulted in premature aging (Rera et al., 2011). Moreover, it is believed that, rapamycin, the drug which increase the lifespan by regulating the protein hemostasis and deregulating nutrient sensing, may also participate in increasing the efficiency of stem cell activity and rejuvenation (Castilho et al., 2009; Chen et al., 2009; Yilmaz et al., 2012).

1.2.9 Altered Intercellular Communication

Intracellular communication alteration in multicellular organisms is a prominent characteristic of aging. "Inflammaging", the pro-inflammatory traits in aging, is one of the consequence of this miscommunication. Inflammaging may rise due to the accumulation of pro-inflammatory damages. In addition, inability of immune cells to eliminate pathogens and senescent cells which have tendency to release proinflammatory cytokines, are some of the factors leading to inflammaging (Salminen et al., 2012).

1.3 Aging and Alzheimer's Disease

Alzheimer's disease is a most common chronic neurodegenerative dementia. the symptoms and severity of which increase over time. However, the rate of changes are various among patients. The prevalence of AD in 2006 was 26.6 million, and this number is expected to quadrupled in 2050 (Brookmeyer et al., 2007). The incidence onset to the AD is increased with age. It is stated that, 12% of the people over 65 carry this disease. Furthermore, this percentage increase to more than 50% in individuals over 85 (Alzheimer's Association, 2011). A study which investigate the microarray-based gene expression changes in aging and AD, found that there are highly statistical overlap between differential expression genes in aging and genes dysregulated in AD (Avramopoulos et al., 2011; Yuan et al., 2012). In addition, genome-wide association studies of 5 different age-related diseases show that they share common age-related pathways (Johnson et al., 2015). Although, aging and AD demonstrate different phenotypes and symptoms, the correlation between them indicates common pathways and mechanisms they may share.

1.4 Network Modeling

Network models consist of biological components and links between them which represent their association. Some of these models are protein–protein interaction, gene

interaction, protein-DNA and metabolic networks. Interaction network tools employ different algorithms to optimally reconstruct biological networks. KeyPathwayMiner is an algorithm which obtain highly connected sub-networks of deregulated genes by employ multiple omics studies. This algorithm apply colony optimization and fixed-parameter algorithms which combine, biological network and multiple omics (Alcaraz et al., 2014). TimeXNet is another algorithm which determine the reliable edges that establish a connection between differentially expressed genes at three initial, intermediate and late time interval by taking weighted interaction network in to account (Patil and Nakai, 2014). SDREM is a network modeling tool which combines two signaling cascade and transcriptional regulation components to examine cellular response to disease. To do this, this algorithm take a upstream proteins which shown related to pathogens and search for signaling cascades which provide interaction of these proteins with downstream transcription factors (Gitter and Bar-Joseph, 2013). SAMNet is an optimization tool that combines two high-throughput data using protein-protein interaction to identify functional groups shared among them(Gosline et al., 2012). Another algorithm, ResponseNet, assert the artificial node (Source) which differentially expressed genes connected to them. Here, this algorithm select optimal connections and reliable nodes taking in to consideration the cost of interactions (Yeger-Lotem et al., 2009). Tied Diffusion Through Interacting Events (TieDIE) uses diffusion model to detect the effect the association of genomic perturbations to transcriptomic changes especially in cancer studies (Paull et al., 2013). The HotNet algorithm also utilize diffusion model to discover modules changed in cancer (Vandin et al., 2011).

1.5 Network Analysis in Aging studies

It is rarely possible that a biological function relies on a single molecule. Instead, biological systems are complex networks that exist through interaction of DNA, RNA and protein components (Barabási and Oltvai, 2004). Studies on biological networks emerge from Albert-László Barabási's finding on scale-free network. He believes that some nodes have higher connections and act as a hub in a system.

With the increase in molecular interaction databases, studies on biological questions with the help of network biology increased. Among these, only few concentrate on aging networks. In 2004, a study examined the connectivity of age-associated proteins and other traits in yeast interaction network. It deduced that senescence related proteins show higher connectivity in a network. In addition, these proteins exhibit high correlation with their degree of pleiotropy which is consistent with antagonistic theory of aging (Promislow, 2004). Following the mentioned study, Ferrarini et al. utilized interaction datasets of *S. cerevisae*, *C. elegans* and *D. melanogaster* which contained physical interactions and published genetic interactions, to create networks. They examined the number of links and local connectivity of age-related proteins. They conclude that, age-related genes act as hubs in a network (Ferrarini et al., 2005).

A "longevity network" was the first time introduced by a study in 2007. In this study, genes which showed association with aging in different species collected and their human orthologs specified. This study searched for the direct interaction of these genes or through first shared neighbors in a human protein-protein interaction. Highly connected proteins (hubs) in this network, including non age-related proteins, reported to be involved in age-related disease (Budovsky et al., 2007).

Another study constructed a "disease-aging network", which shows the interaction between age-related genes and disease-related genes. This network demonstrated that the average closeness centrality of aging genes is much higher than disease-associated genes. In addition, genes which associated with aging establish a connection between disease, especially age-related ones (Wang et al., 2009). A similar study examined the common modules which were conserved in a co-expression network of aging and AD. Energy metabolism of mitochondria and synaptic plasticity were two common functional groups enriched in this study (Miller et al., 2008).

1.6 Research Objective

As discussed earlier, aging is an accumulation of time-dependent deleterious changes in biological processes, which leads to the physiological disruption; hence raise the

prevalence of age-related diseases such as Alzheimer's disease. Therefore, understanding the mechanisms of aging is important in discovering the molecular mechanisms of these disease. I believe high correlation of differential expressions between aging and AD indicates the common pathways these two processes may share. Its known that aging and AD are highly complex and gene expression levels are heterogenous. Therefore, this emphasize the importance of analysing these genes and their interactions in a transcriptomic network.

However, some weaknesses on microarray data measurments such as difficulties in detection of low light signals, inability of perfect match bindings or inefficiency of platforms limit the informations collect from data-set. In addition, post-translational modification may affect expression changes and does not shown in micoarray.

Here, I have tested Omics Integrator Software (Tuncbag et al., 2016) to reconstruct the optimal aging ad AD networks. This software employ prize-collecting Steiner forest problem to achieve a network which contains high reliable genes and confidence interactions. This algorithm allows genes to interact directly or throughout intermediate or steiner nodes. I believe, this could eliminate some biases in microarray measurments which mentioned earlier. In addition, I believe functional analysis of clusters of aging and AD networks would helps to obtain Gene Ontology results and pathways shared between them.

CHAPTER 2

MATERIALS AND METHODS

2.1 Datasets

In order to construct the brain aging and Alzheimer's disease networks I used microarray based gene expression datasets. All of these data were retrieved from NCBI's Gene Expression Omnibus (GEO) data repository and had been generated on the Affymetrix platform (Edgar et al., 2002). The priority of extraction for Affymetrix datasets is to download "CEL" files (raw data), which harbor light intensity measurements (Gautier et al., 2004). Pre-processed files, called a "series matrix files" are also available at NCBI GEO. This form of file is background corrected and normalized by the authors. However, I chose to start my analysis from raw data whenever possible.

In dataset selection I took a number of features into consideration. One was high sample size, which is critical for achieving statistically significant results. In addition to the sample size, the age range was taken into consideration, such that I tried to maximize the interval between developmental process termination (20 years in human) and old age, which is important in studies of aging. Three different datasets, two human and one mice, which in total comprise 4 different brain regions were downloaded in order to be used in this study, shown in **Table 2.1**. In this table data belonging to each brain region in a dataset is shown separately, and I refer to each of these as a "dataset" in further chapters. Note that the sampled brain region in the mice study is not specified in related article (Jonker et al., 2013).

All of the datasets I chose had unprocessed "CEL" files available. I processed these

raw datasets with the help of the R "affy" package, to analyze expression levels (Gautier et al., 2004). I would like to note that newly designed Affymetrix oligonucleotide array platforms are not supported by this package and the R "oligo" package should be used (Carvalho and Irizarry, 2010).

For Alzheimer's disease, I used 2 datasets, comprising 3 brain regions in total, presented in **Table 2.2**. Here, similar to **Table 2.1**, each region is shown separately and I call each of these a "dataset". All the datasets contain high number of non-dementia and AD samples. Despite the availability of raw data for these datasets, I downloaded "series matrix files" from NCBI GEO. The reason is that the platform used in these two studies, "Rosetta/Merck Human 44k", is not supported by "affy", "oligo", or any other freely available R package, to my knowledge. I note that this platform's preprocessed series matrix files contained negative values. Negative values may arise due to the higher expression of background probes relative to perfect match ones.

The data from glioblastoma multiforme (GBM, the most aggressive type of brain cancer) patients deposited in TCGA (the Cancer Genome Atlas) were also used in this project. GBM data was retrieved from the Genomic Data Commons (GDC) data portal (https://cancergenome.nih.gov/). Different categories of data such as DNA methylation, DNA sequencing, transcriptome profiling and copy number variation are available in this portal. RNA-Seq expression files with FPKM (Fragments Per Kilobase Million) units were downloaded (**Table 2.3**). Expression of each gene in this dataset represents the amount of reads mapped to a gene's annotated location. The problem with this dataset was the low number of control samples (5 controls in 161 samples). Therefore, statistically appropriate estimation of differential expression is not possible.

2.2 Preprocessing of Datasets

As I explained in **section 2.1**, my priority is to extract data from the raw data. The advantage of using unprocessed data is that, the same normalization can be apply

Table 2.1: Aging datasets. The column "Dataset ID" represents the name of first author of the study and the abbreviation of the brain region involved. "Yrs" in "Age Range" column represent years of age.

Dataset ID	**Organism**	**Brain Region**	**Sample Size**	**Age Range**	**Platform**	**GEO accession number**
Berchtold_HIP	*Homo sSapiens*	Hippocampus	43	20-99 Yrs	HG-U133_Plus_2	GSE11882
Berchtold_SFG	*Homo sSapiens*	Superier frontal Gyrus	48	20-99 Yrs	HG-U133_Plus_2	GSE11882
Berchtold_PCG	*Homo sSapiens*	PostCentral Gyrus	43	20-99 Yrs	HG-U133_Plus_2	GSE11882
Lu_FC	*Homo sSapiens*	Frontal Cortex	30	26-106 yrs	HG_U95Av2	GSE1572
Jonker_Brain	*Mus musculus*	Brain	18	13-130 weeks	Mouse430_2	GSE34378

Table 2.2: Alzheimer's Disease datasets. The column "Dataset ID" represents the name of first author of a study and the abbreviation of the brain region. In the "Conditions" column, ND is non-dementia and AD is Alzheimer's disease.

Dataset ID	**Organism**	**Brain Region**	**Sample Size**	**Condition**	**Platform**	**GEO accession number**
Narayanan_PFC	*Homo sSapiens*	PFC	467	157 ND and 310 AD	Rosetta/Merck	GSE33000
Zhang_CR	*Homo sSapiens*	Cerebellum	230	101 ND and 129 AD	Rosetta/Merck	GSE44772
Zhang_VC	*Homo sSapiens*	Visual Cortex	230	101 ND and 129 AD	Rosetta/Merck	GSE44772
Zhang_PFC	*Homo sSapiens*	PFC	230	101 ND and 129 AD	Rosetta/Merck	GSE44772

Table 2.3: Glioblastoma Multiforme dataset. The column “Dataset ID” represents the name of the project and cancer type. In the “Conditions” column, “GBM” is Glioblastoma Multiforme.

Dataset ID	**Organism**	**Brain Region**	**Sample Size**	**Condition**	**Platform**	**GEO accession number**
TCGA_Glioblastoma	*Homo sSapiens*	-	156	156 GBM	Illumina	TCGA project

generically to all datasets, and therefore reduce possible bias in meta-analysis. In data generated with Affymetrix, CEL files contain the light intensity of probes of a microarray. Converting these signal intensities to the relative gene expression levels for each probe can be performed using the free R packages "affy" or "oligo". For newly designed affymetrix arrays the "affy" library is dysfunctional, and the "oligo" library should be used. These libraries can be accessed through Bioconductor open source software project (https://www.bioconductor.org/) (Gentleman et al., 2004).

All of the raw data I investigated in the aging study, Berchtold_HIP, Berchtold_SFG, Berchtold_PCG, Lu_FC and Jonker_Brain were supported by the "affy" package. This package's function "ReadAffy", takes the directory of files, the name of the CEL files which will be read, the Chip Definition File (CDF) name and other possible parameters. The CDF file contains the layout information of probes on a chip. The "ReadAffy" function can detect automatically the platform's related CDF. The "Cdfname" argument in the "ReadAffy" package can also be used to specify the name of an alternative CDF library. I used a costum CDF from the Brainarray database which I will explain in **section 2.2.3** in detail (http://brainarray.mbni.med.umich.edu/Brainarray/default.asp). During data processing, the affy library's "expresso" function is used to apply normalizations such as RMA background correction method, summarization across probes, and quantile normalization across samples, which I will explain in details in next sections.

The raw data of AD datasets exist, however they are not in CEL file format; an R package supporting this platform does not exist, and no freely available software is available to my knowledge. Therefore, I used pre-processed NCBI GEO "series matrix files" for the datasets Narayanan_PFC, Zhang_CR, Zhang_PFC and Zhang_VC. Quantile normalization was the only processing step I applied to these datasets.

The Glioblastoma Multiforme dataset contains FPKM values of each gene. This file has a high number of zero values and genes which are not expressed in any of the samples. Following the removal of these genes, I applied quantile normalization.

2.2.1 RMA Background Correction

Microarray is a technology to detect relative expression levels of multiple genes on a single chip. Each microscopic spot on a chip contains a DNA oligomer, known as a probe (oligo) (Gerhold et al., 1999). On a standard commercial chip type, each spot at a specific position has the identical sequence across chips. The expression level of a gene is estimated using data from the collection of different oligonucleotide probes designed to measure that gene's cDNA, derived from mRNA. Each probe is either complementary to a gene's cDNA, known as a perfect match (PM) probe, or there exists a substitution of one nucleotide in the probe sequence which prevents perfect binding. These latter probes are used to detect background (non-specific) hybridization and noise in expression, and are known as mismatch (MM) probes (Hubbell et al., 2002). In the experimental procedure, the intensity of each hybridization signal between mRNA/cDNA and the fluorescent probe at a specific spot is measured and provided in a CEL file (Schena et al., 1995; Gautier et al., 2004).

The noise and non-specific bindings detected on the microarray chip are subtracted in the RMA background correction method. In addition, this method, prevent the interfusion of neighbors signals (Parmigiani et al., 2003). Later, intensity value of probes that arise from probes designed for the same gene across the array are combined, and this average represents the expression level of the gene. The RMA algorithm's advantage over other background correction algorithms is its detection ability of slight expression signals (Irizarry et al., 2003).

The affy package's "expresso" function performs RMA background correction. Distribution of the expression values obtained in this method is right-skewed due to the high amount of low expressed genes and low amount of highly expressed ones. Therefore, "expresso" transforms the data to log2 base, which provides lower variance for large values, helps the data to fit a normal distribution, and reduces the dependency between mean and variance. After log-transformation the data is less influenced by a few highly expressed genes, easier to visualise, and also appropriate for analysis with parametric statistical methods, many of which assume a normal distribution and equal variance (Whitlock and Schluter, 2009).

2.2.2 Quantile Normalization

A large sample size is important in transcriptome data analysis. Quantile normalization is used to eliminate noise among samples that arises due to technical issues and to unify the distributions in multi-sample data. This method assumes that sample variation is due to technical noise. However, we should be aware that it may eliminate interesting biological variations (Hicks and Irizarry, 2014).

As I previously mentioned, the affy package's "expresso" function employs quantile normalization in the raw data normalization process. However, in series matrix files I used the "preprocessCore" package (which is also part of the affy project).

2.2.3 ID Conversion

In order to compare different datasets and continue further analysis, probe-set IDs should be converted into a common gene identifier. The package "biomaRt", the interface to the Ensembl Biomart database, is the most preferred ID conversion tool (Durinck et al., 2005). However, the data retrieved from this library does not consider the situation that each probe-set of a microarray platform may correspond to more than one Ensemble ID, and each Ensembl gene ID may correspond to more than one probe-set.

To deal with the mentioned problem, in analyzing raw data files I used custom CDFs prepared by the Brainarray project. The filtration procedure designed by this group are; (a) Blast alignments of probe sequences against cDNA and EST sequences must be perfect matches. (b) Each probe should represent one uniGene cluster and map to the same genomic location. (c) Probes which belong to the same cDNA should align to the same genomic location and direction. (d) Each probe-set must have at least three different probes (Dai et al., 2005). The above steps lead to one-to-one conversion of probe-sets to Ensembl gene IDs. Additionally, I filtered and converted mice Ensembl genes to their one-to-one orthologous human gene IDs with the help of the "biomaRt" package.

The above mentioned solution is applicable when raw data exist and the Affymetrix platform-related Brainarray custom CDF file is available. The other solution for probe-set to gene ID conversion for files that do not match the above conditions is (1) to get rid of probe-sets representing more than one Ensembl gene, and (2) take the average of expression level per sample across probe-sets corresponding same Ensembl gene ID. In some studies, instead of taking the average, the maximum expression level is also preferred. The common aim of microarray platforms is to choose sequences that perfectly match with the targeted transcript and show lowest similarity to the rest of whole genome. However, at least in Affymetrix microarrays, different probe-sets are usually designed to measure expression from different transcripts of the same gene (Liu et al., 2010). It is known that alternative splicing changes during aging and age-related neurodegenerative disease (Mazin et al., 2013). Therefore, using the probe-set with the maximum expression level per gene to represent that gene's expression level could lead to elimination of heterogeneous transcript effect in gene expression level measurements. Therefore, using the average expression level is a more reliable approach. In addition, I find out that the Brainarray custom CDF file shows higher correlation with the datasets created using the average method rather than the maximum method (data not shown). I believe, taking average of probes, following probe filtration in brainarray, is the reason of this consistency.

The annotation file of the "Rosetta/Merck Human 44k" platform used for the AD datasets was provided in the NCBI's GEO data repository. This file contains probe-set IDs, mapped Entrez gene IDs, and other annotations. Therefore, firstly I converted probe-set IDs to Entrez gene IDs and then to Ensembl gene IDs with the help of "biomaRt" package, and apply average method at each step.

2.3 PCA

Principle component analysis is an algorithm that is used to summarise and visualise variation in a dataset. This algorithm is used when there are wide range of variables and visualization of samples' similarity and differences is difficult. It reduces

the multi-dimensionality of the dataset by assigning linear combination of variables, known as principal components (Ringnér, 2008). The first principal component (PC1) is a line in a multi-dimensional space that explains the largest variation.

The two main principal components (PC1 and PC2) explain the largest among of variance in a dataset and are usually used to check the clustering of samples and outliers. In addition, PC3 and PC4 also frequently checked, considering the proportion variance they explain. Outlier samples could raise noise in transcriptome analysis. Following removal of outliers I check whether the number of genes showing significant differential expression decrease or increase upon removal of the outlier, with the expectation that removing an outlier (which introduces technical noise) should improve the differential expression signal. Preprocessing of dataset and quality control are repeated following the removal of each outlier. Here, principal component analysis was conducted with the help of the R function "prcomp".

In addition to principal component analysis, two alternative newly developed outlier finder methods exist: robustPCA and bagplot. Robust principal component analysis separates the data into two matrices, sparse and low-rank, by principal component pursuit approach. The sparse matrix which deposit noise in a data is employed to find the variations among samples (Candès et al., 2011). Meanwhile, the bagplot method is used for detecting the data variation in a bivariate boxplot (Rousseeuw et al., 1999). I have examined these two methods. However, due to inconsistency of these methods' outcomes with PCA (data not shown) and more efficient interpretation of classic PCA results I decided not to include them in this project.

2.4 Differential Expression

To evaluate possible monotonic relationships between gene expression and age and identify potential age-related genes, I used the non-parametric Spearman correlation rank test. In contrast to Pearson correlation, this method calculates the relationship by ranking the data, and is robust to variations in the data such as outliers (Hauke and Kossowski, 2011). In molecular aging studies it is generally assumed that gene

expression change with age is gradual and linear. Therefore I did not search for non-linear associations (which would not be detected by pearson correlation).

For each gene I applied Spearman correlation between age and gene expression. Two results, the p-value and the correlation coefficient inform about the significance, and the degree and direction of association, respectively. The correlation coefficient (ρ) can range from -1 to 1. Minus 1 represent strong negative relationship and plus 1 represent positive relationship. Omics Integrator Software, which will be explained in further sections, does not accept negative values. Therefore, I used differential expression information by taking the absolute value of these correlation coefficients. This forced us to analyse increase and decrease gene expressions at the same time. Here, I am assuming that decrease and increase changes can affect same pathways.

However, in the Alzheimer datasets, there are two conditions, control and AD. Therefore, in order to find differential expression, I could also use the non-parametric Mann-Whitney U Test for testing difference in medians. However, to be consistent with the aging results, I preferred to use correlation coefficients in all the network analyses, and therefore I applied Spearman correlation test by defining two different stable numbers to condition variables. I note that the p-values of the two methods, Spearman and Mann-Whitney-U tests are close to each other. The "wilcox.test" function in R conducts Mann-Whitney-U test.

2.5 Multiple testing correction

Evaluating p-value results, obtained from simultaneous statistical tests separately applied to a large number of genes, is not enough for identifying significant genes. The reason is, there is a possibility that some of these values randomly have nominally significant p-values, but are not reproducible. Therefore, I am expecting to have high number of type I errors (false positives) when performing large-scale statistical comparisons (Benjamini and Hochberg, 1995). The "false discovery rate" (FDR) approach can account for this type I error inflation.

There are multiple FDR methods and they are provided by the R function "p.adjust". I used one of the most powerful FDR methods, "Benjamini Hochberg", on p-values (Benjamini and Hochberg, 1995). The results of multiple testing correction are called q-values. I applied a cutoff 0.1 to q-values to eliminate these false rates (Verhoeven et al., 2005).

I should mention that, although FDR methods reduce type I error, they also rise type II error and this may lead to the failing to detect biologically important effects.

2.6 Correlation between datasets

I have selected 5 aging and 4 Alzheimer's disease gene expression datasets in this project. I am expecting that datasets which derive from common tissues or involve the same biological patterns should show high correlation. In order to find correlations between datasets, I applied pairwise Spearman correlation to correlation coefficient of shared genes between two datasets without applying multiple testing correction.

2.7 Omics Integrator Software

I used the Forest module of the Omics Integrator software (Tuncbag et al., 2016) to reconstruct optimal network for each dataset. Forest module solves the prize-collecting Steiner forest problem to integrate multiple data in a network context. The aim of this algorithm is to optimally connect selected genes, called terminals, by using a template interactome. Each interaction in the interactome is weighted with its confidence score. Each terminal has a given prize of their correlation coefficient. This value represent the direction and strength of association between expression and condition. Forest module searches for the optimal network either by linking the terminals directly or through intermediate nodes, called Steiner nodes. MI-score is the method used to calculate the confidence score of interactions in a network. This method emphasize the significance of interaction by a measure based on the number of publications reporting their association (Villaveces et al., 2015). On the other hand, subtracting

confidence scores from 1 represents the cost of interaction.

$$Cost = 1 - ConfidenceScore \tag{2.1}$$

Forest algorithm harbors other parameters to reconstruct biologically meaningful networks. PPI is a combination of interactions frequently used in the literature. Some proteins are critically important in biological studies and there are the focus of many studies. Therefore, some proteins' interactions are discovered more than others. To illustrate, ubiquitin C (UBC) shows approximately 7407 connections (degree) in a irefindex network. This amount of degree could lead to bias in the forest algorithm. Because each terminal would try to connect to the other terminal using UBC as a shortcut. The parameter μ, is a multiplier which gives a node a penalty for each edge addition. The formula of prize calculation and negative weighting of each terminal is as below:

$$p'(v) = \beta.p(v) - \mu.degree(v) \tag{2.2}$$

p'(v) is a new prize of the terminals and p(v) is the initial prize. The value of β controls the size of the final network. The larger β values force the network to include more terminals which may lead to inclusion of low confidence edges.

Forest module minimizes the objective function in **Eq. 2.3** where it minimizes the total prizes not included and the total cost of edges in the final network.

$$f'(F) = \sum_{v \notin V_F} p'(v) + \sum_{e \in E_F} c(e) + \omega.\kappa \tag{2.3}$$

V_F and E_F represent the set of vertices and edges in a network F respectively. c(e) is the cost of edges and the κ is the number of sub-trees in the forest, F. In a nutshell, this formula calculates the sum of prizes of the terminals which are not included in the final network, cost of interactions and the number of sub-trees. ω is the parameter to determine the number of sub-trees in a reconstructed network. To do this, an artificial node is connected to subset of all nodes, with the edge cost of ω and after optimization is complete that artificial node is removed from the network to collapse it into multiple subtrees (Tuncbag et al., 2013). The prize-collecting Steiner tree problem is solved with the message-passing algorithm implemented in msgsteiner tool (Tuncbag et al.,

2016). Schematic example of forest algorithm application is given in **Figure 2.1**. In this figure, the algorithm does not permit the small prize and costly edges to get into network.

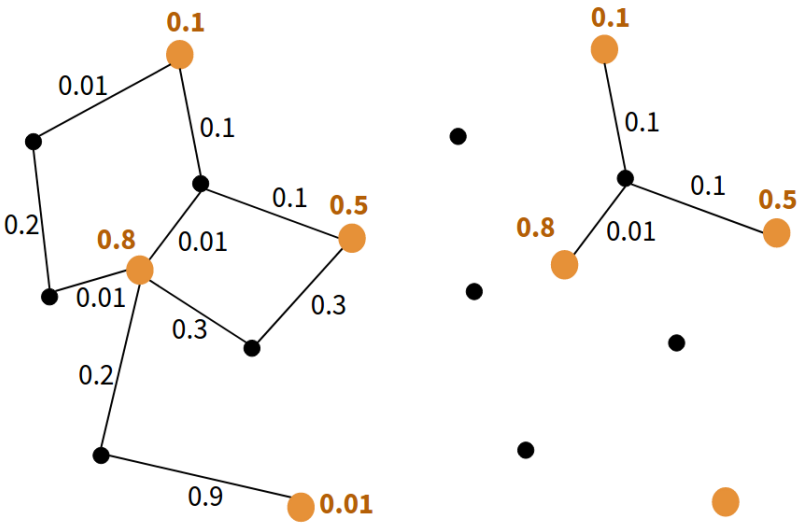

Figure 2.1: A toy example of the prize collecting Steiner tree algorithm. The network on the left shows terminal nodes in orange in a protein-protein interaction. Forest algorithm tries to link terminals optimally by using a template network as shown in the left panel and reconstructs the final network shown in the right panel. Nodes and node labels colored orange represent terminals and their correlation coefficient, respectively. Edge labels colored black represent edge costs.

Finding optimal parameter values needs a tuning step. The factors determine the outcome of forest algorithm are the number of terminals (selected genes) covered in the final network, distribution of the correlation coefficients, degree distribution of the nodes in the final network and edges costs. Therefore, I planned to apply forest algorithm from the OmicsIntegrator package to top 800 significant genes in each transcriptome dataset with different combinations of μ, ω and β, and check the network features. The aim of selecting this number of genes is to provide amount of genes which forest algorithm can handle, and achieve biologically meaningful results. In this parameter tuning, I searched for a network which contains highest number of terminals. Among those that contain the same terminal count, I selected the network with highest number of nodes.

2.8 Protein-Protein Interaction Network

OmicsIntegrator package also included irefindex protein-protein interaction dataset (Turner et al., 2010). I converted the gene annotations from gene symbol to the Ensembl gene ID. Here, I removed Entrez gene IDs represent more than one Ensembl gene ID and vice versa.

In the previous section I mentioned that, genes have high degree counts create bias in a network and I include the μ parameter to exclude this problem. However, ubiquitin C contains extraordinary number of degrees and the mentioned parameter could not deal with it. Therefore, I removed UBC node from the PPI. This ejection could increase type II error, but eliminate possible bias in a network.

2.9 Network Clustering

Hairball-like structure of the networks does not provide too much information about biological characteristics of it. I am assuming that biologically-related genes which share common gene ontologies or pathways are highly connected in a network but this has to be shown explicitly.

In this project I used the "louvain modularity" algorithm. This algorithm takes a node and searches for neighbors which maximize modularity. Newly formed community is represented as node in a network and the algorithm searches for neighbors again. This procedure is continued recursively until it fails to increase modularity (Blondel et al., 2008). I removed clusters with lower than 20 nodes to have better statistical results in enrichment analysis.

2.10 Enrichment analysis of the network clusters

I applied Gene Ontology (GO) and KEGG pathway enrichment analysis to clusters obtained from each network. Cellular component (CC), molecular function (MF) and

biological process (BP) are three domains of this ontology. Other ontology terms are hierarchically branched under these domains. BP terms contain molecular activities which have defined start and end (Ashburner et al., 2000). On the other hand, KEGG pathway provides the information about the connections of gene products (Kanehisa and Goto, 2000).

Enrichment analysis calculates the chance of selected genes in a functional group to background. It is clear that, to analyse each network individually we should take all genes in a gene interaction network as a background. However, in this meta-analysis study I am searching for common patterns among networks. Therefore, to determine background, from gene interaction network, I discarded union of genes in aging and AD forests separately. **Figure 2.2** exhibit schematic explanation of background selection.

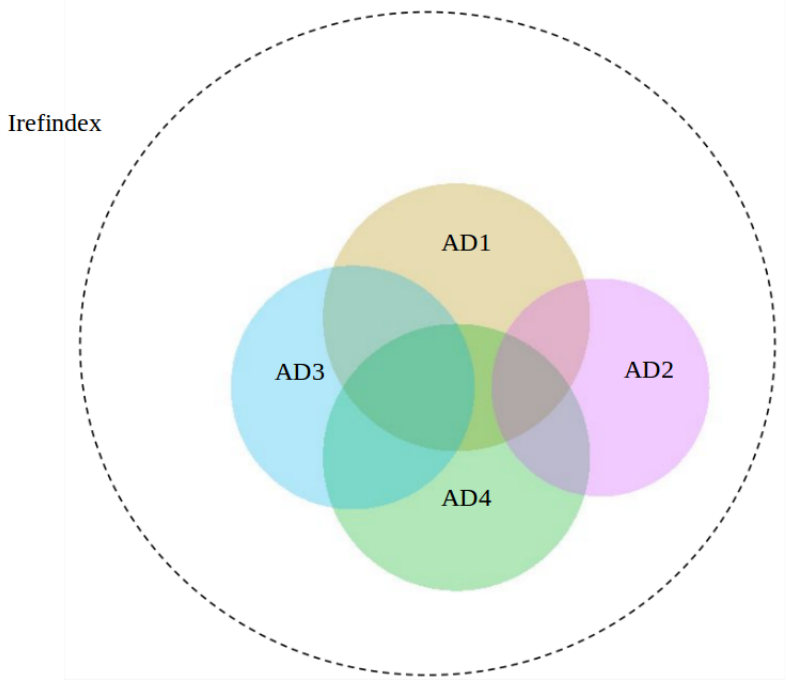

Figure 2.2: Schematic example of background selection. Irefindex is a gene interaction network (big circle) and colorful circles are sample AD reconstructed networks. Background is a collection of genes that fall into the white region.

I applied Biological Process Gene Ontology analysis with the help of R package

“topGO”. This package provides various algorithms. To illustrate, topGO’s default algorithm “weight01” applies enrichment analysis from bottom to top and each time removes the child term genes from parent term. In addition, it takes the position of terms in a hierarchy into account. In this project I applied Fisher’s exact test with classic algorithm which examine GO terms independently. Also, I eliminate terms contain lower than 10 genes.

For KEGG pathway enrichment analysis I did not use any package. The reason is, interface packages of the DAVID database accept limited number of genes. This made a problem as I used custom background. Therefore, after downloading pathway and gene annotation data from the KEGG database resource, I applied EASE Score, a modified Fisher’s exact *p*-value (Aoki and Kanehisa, 2005). This method decreases type I error in outcomes.

2.11 Common Edges

I searched for common interactions among networks and applied Gene Ontology and pathway enrichment analysis on genes that provide these interactions. I also checked whether these genes are included in the AnAge (http: //genomics.senescence.info/species/) database or not. The mentioned database contains genes show relatedness to longevity researches.

2.12 Permutation

Possible bias in each step of the analysis could leads to different consequences. To examine significance of my results, three kinds of permutation test was applied. "Noisy Edges" is a function in Forest algorithm command (Tuncbag et al., 2016). The value given to this function determines how many times this algorithm adds noise to the edges. Another function, "Shuffle Prizes", specifies the number of times the prizes shuffle in a network. In addition to these two, I randomly shuffled ages in the gene expression data. This changes differential expression information and the consequences

of forest algorithm. These permutations were applied 100 times to all datasets with using forest algorithm parameters obtained from parameter tuning for each dataset. However, some of the permutations caused optimization problems and which led to empty networks. Therefore, for each permutation type, I calculated the lowest number of non-empty networks among datasets as a permutation count (n). Subsequent to clustering the data, in order to apply enrichment analysis I took n^{th} non-empty result from each datasets and selected the background as I explained in **section 2.10**.

CHAPTER 3

RESULTS

3.1 Differential expression in each dataset

In this study, I used 9 different gene expression datasets from various brain regions, conditions and datasets, including aging and Alzheimer's Disease (**Tables 2.1** and **Table 2.2**). As I explained in Chapter 2, my priority was to use raw microarray data (light intensity measurements) without any preprocessing procedure applied to them. However, four AD datasets, Narayanan_PFC, Zhang_CR, Zhang_VC and Zhang_PFC do not contain raw data files which supported by freely available R package. Following pre-processing, background correction, normalization and ID conversion of data, I checked the variation in a datasets with the help of PCA. PCAs of Berchtold_PCG and Narayanan_PFC are given in **Figure 3.1** and **Figure 3.2** respectively.

Biologically close samples such as similar ages or AD patients are expected to cluster together along principal component analysis trajectories. Also the cluster of samples in a plot might be the sign of a batch effect (technical similarity, such as sample processing day). Apart from this, some samples demonstrate different gene expression patterns due to other biological problems. These outlier samples could raise noise in transcriptome analysis and could reflect themselves in a PCA plot. The red dot in **Figure 3.1**, which I accepted as an outlier, illustrates this idea.

Following outlier removals, I applied differential expression test on the transcriptome datasets to identify the effect of aging or AD. Here, I used Spearman correlation rank

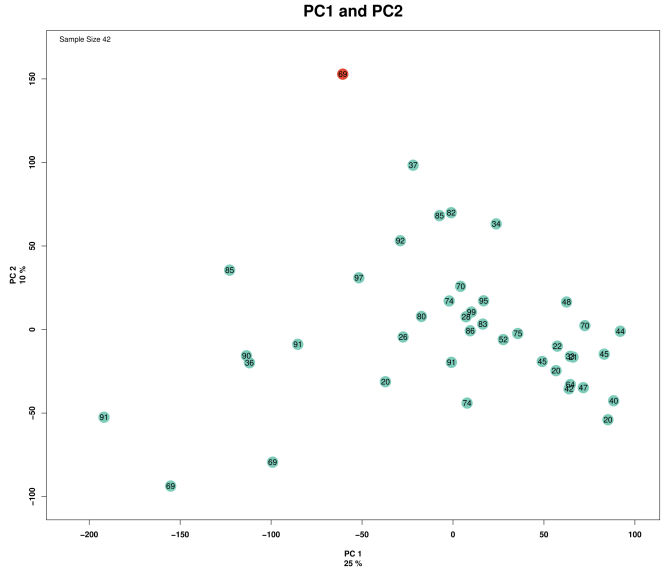

Figure 3.1: PCA analysis of Berchtold_PCG dataset. The plot shows principal component 1 (PC1) and PC2 results. Each dot on a plot represent the samples and their ages. The percentages in each axes indicate proportion of variance of components.

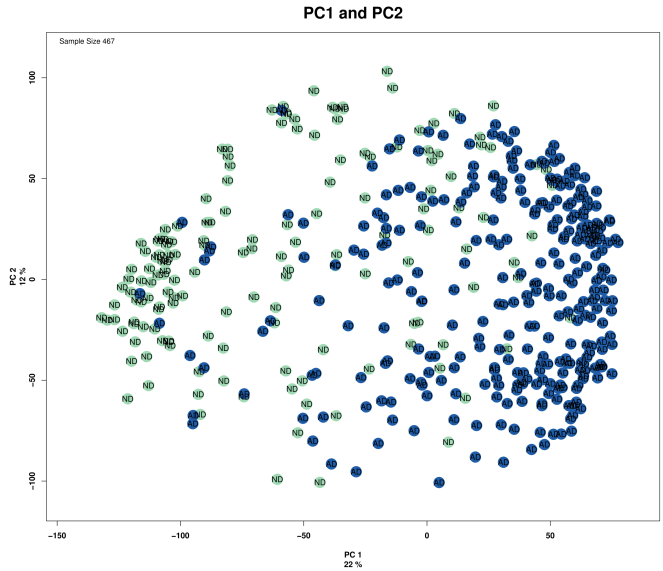

Figure 3.2: PCA analysis of Narayanan_PFC dataset. The plot shows principal component 1 (PC1) and PC2 result. Each dot on a plot represent the samples and their conditions. “AD” stands for Alzheimer’s disease and “ND” stands for non-demented. The percentages in each axes indicate proportion of variance of components.

test to find correlation between gene expression and age for aging datasets. For AD datasets, I also applied Spearman correlation between gene expression of control and AD individuals, using binary coding for disease status. In addition to *p*-value, this method provides correlation coefficient which express the strength and direction of the relationship. In order to eliminate false positive, I also performed multiple testing correction. Then, I chose the most significant 800 genes as the terminal set to use in network analysis. As shown in **Figure 3.3** the total number of measured genes varies among datasets. In addition, the number of differentially expressed genes are much lower in aging datasets compared to AD datasets. However, AD datasets show high numbers of significantly differential expressed genes relative to the number of total genes. This could demonstrate the heterogeneity of genes in AD. In addition, this could rise due to the high number of samples which control spearman correlation results. Jonker_Brain, which is a mouse dataset, contained the low number of differentially expressed genes. The reason could be the poor quality of this dataset.

3.2 Consistency among datasets

I am expecting that the same type of biological samples, such as same tissue or conditions, should share close expression patterns. To identify consistency among datasets, I tested pairwise Spearman correlation between two datasets across all their common genes' Spearman correlation coefficients (between expression and age or AD). This determines the power and direction of associations. The results sketched in **Figure 3.4** with the help of "corrplot" function in R.

All of the datasets are correlated significantly (p-value<0.05). Correlation coefficients are distributed between 0.07 to 0.98, with a standard deviation of 0.23. It is clear that all of AD samples are tightly clustered together. Jonker_Brain is a mouse dataset and appears as an outgroup in hierarchical clustering (data not shown). This could be due to the species difference of this dataset. There are also a cluster between aging and Alzheimer's disease samples. In addition, I applied PCA to detect biological and

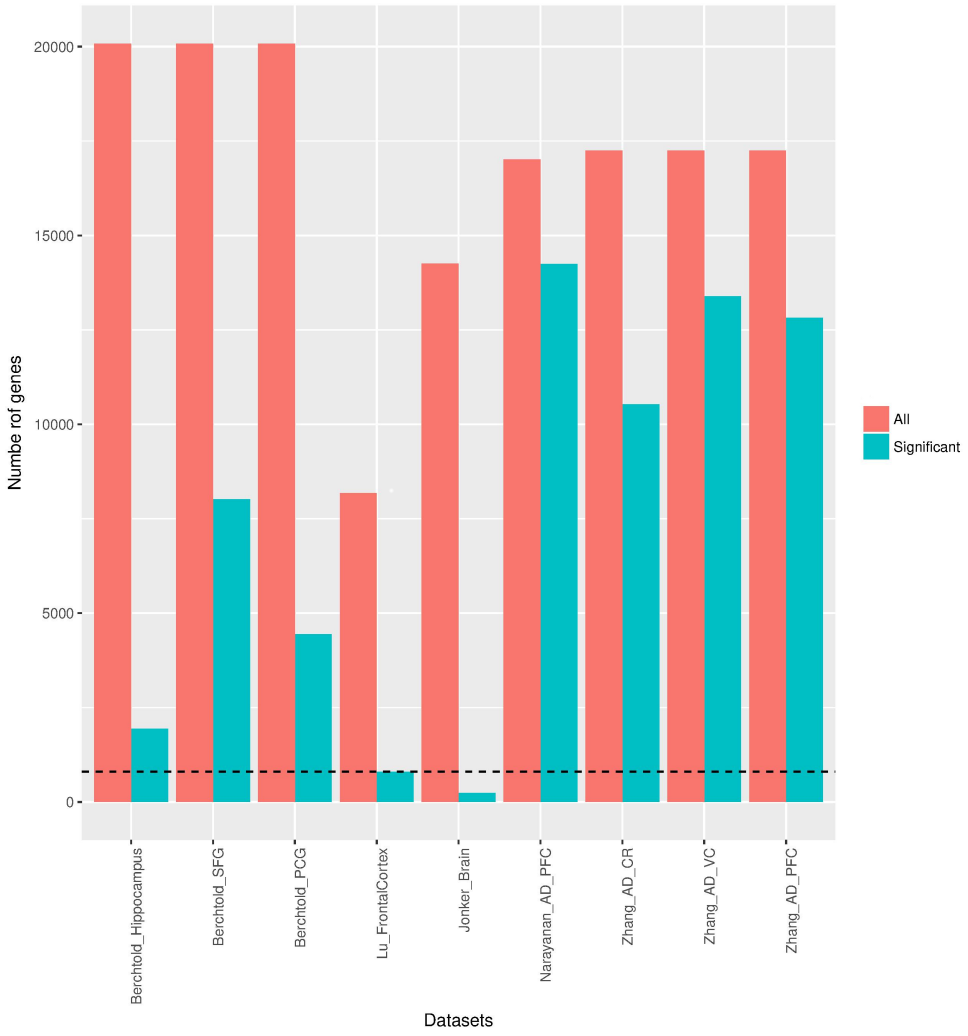

Figure 3.3: Number of genes affected by aging or AD in each dataset. Red bars represent all genes measured, and green bars represent genes showing significant differential expression with respect to aging or AD. Black dash line represent 800 genes chosen for network analysis.

technical variations among datasets. Here, I chose genes which are common among all datasets (n=6448). **Figure 3.5** demonstrates the PCA result. Here, it is clear that except for Jonker_Brain, aging datasets are clustered together. AD datasets are also close to each other.

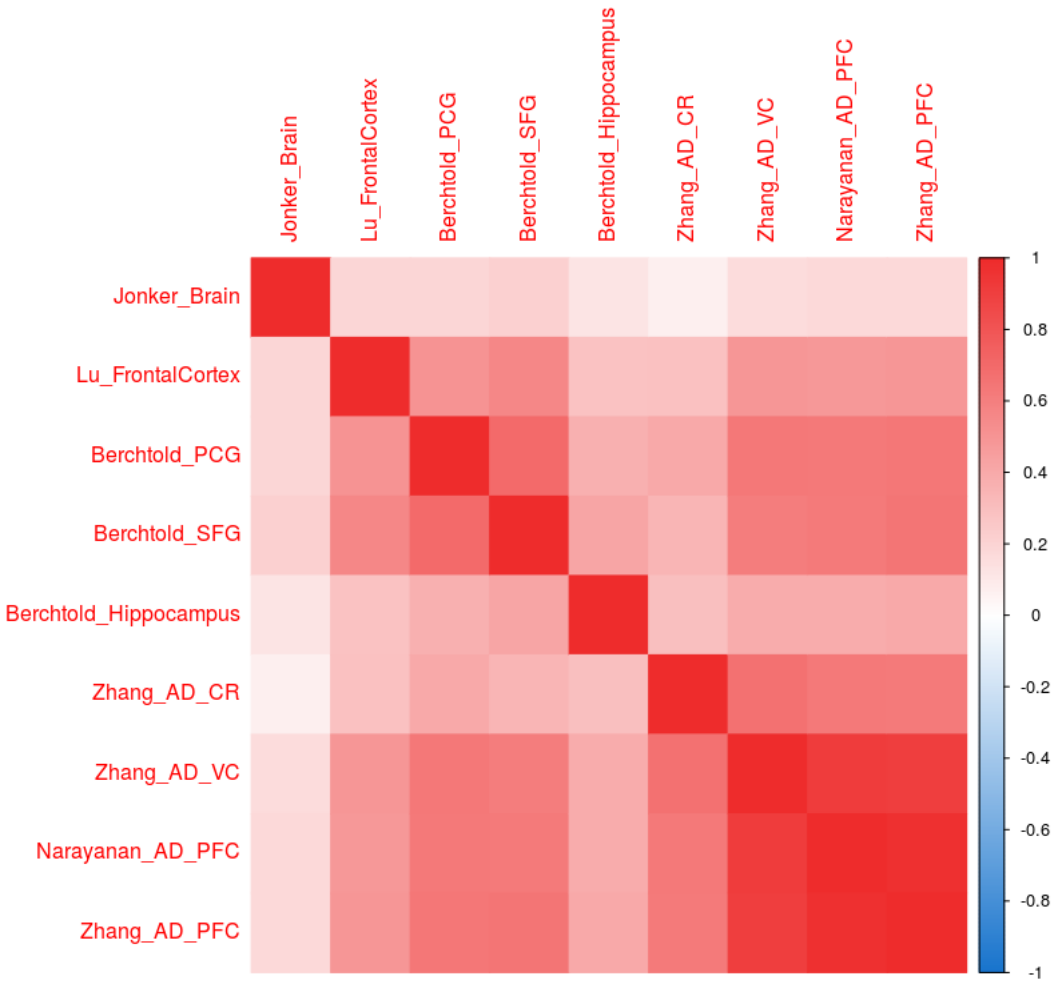

Figure 3.4: Consistency among datasets in gene expression changes during aging and/or AD. Dark red represents highest positive correlation (in age/AD vs. expression correlation coefficients between two datasets across all common genes) and dark blue represents highest negative correlation.

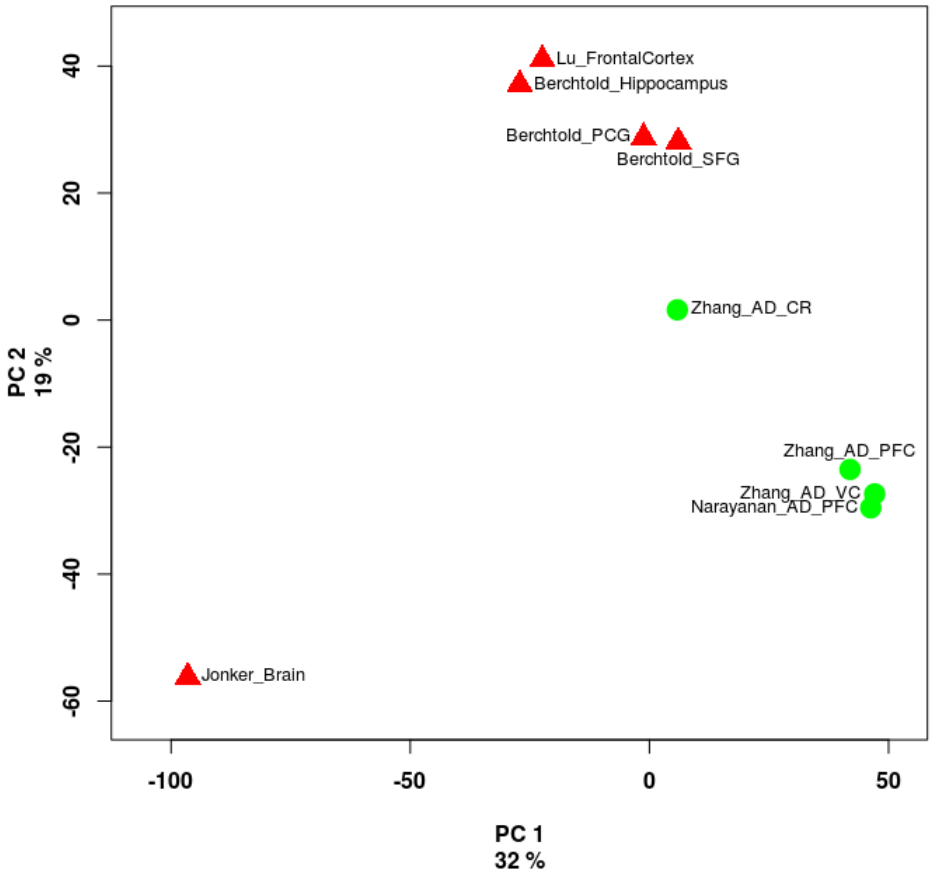

Figure 3.5: Principal component analysis of shared genes' ρ values. Red triangles represent aging and green circles represent AD datasets.

3.3 Forest Algorithm

As I explained in Chapter 2 **section 2.7**, the forest algorithm optimally connects selected genes (Terminal) in a gene interaction network directly or through intermediate nodes. In addition, parameters in forest control the optimization to yield biologically reasonable networks. Before this, it is crucial to explore characteristics of an initial protein-protein interaction network obtained from iRefWeb. Here, I generated the degree distribution of iRefWeb where UBC is removed from this analysis, because it binds almost all nodes in the network.

Degree distribution of biological networks exhibit exponential distribution. This idea was firstly claimed by Albert-László Barabási that biological networks are in scale-free rather than a complex one (Barabasi and Albert, 1999). This means that there are few protein numbers with high degree and high number of proteins with low degrees. I checked the degree distribution and other features of the iRefWeb network. **Figure 3.6** shows that the degree distribution of this network follows a power law.

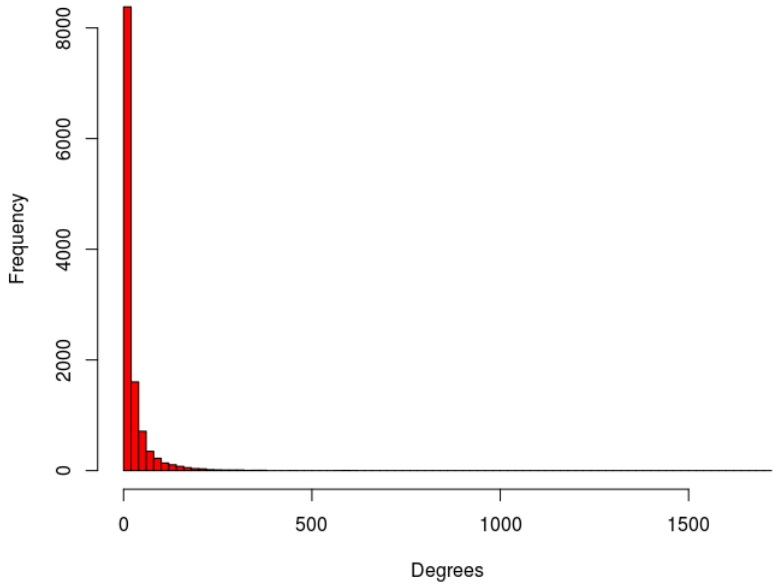

Figure 3.6: Degree distribution of iRefWeb. Protein IDs were converted to gene ID and UBC removed from the network.

Forest parameters μ, β and ω control the degree number-related penalty, the amount of terminals to preserve, and the number of sub-trees in a network, respectively. I tested different combinations of β and ω values in a forest using a fixed μ value 0.01. A sample for parameter tuning is given in **Figure 3.7**. Then, I selected the network which contains the highest number of terminals and nodes. Characteristics of these networks are given in **Table 3.1**.

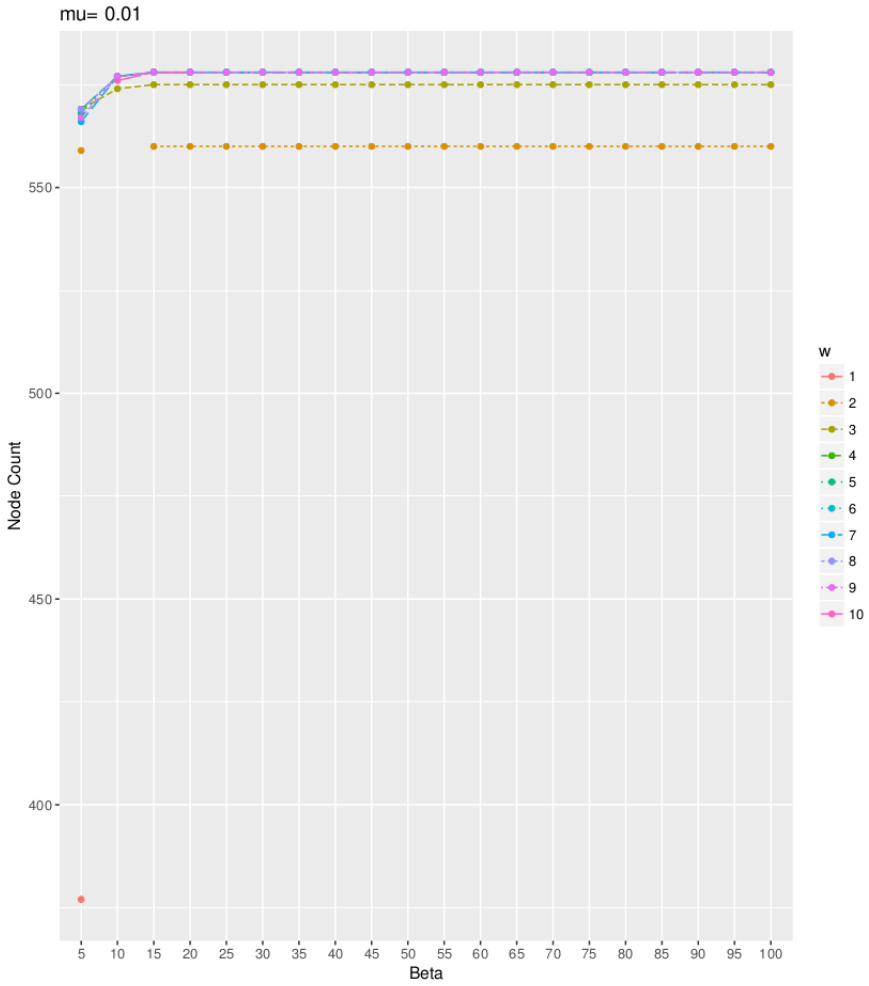

Figure 3.7: Parameter tuning of forest in Berchtold_PCG dataset. Each dot in a graph represent a reconstructed network. X-axis contains β values and y-axis contains terminal nodes count. Each line with different color represent ω values. Parameter μ is constant.

Networks are mostly larger than 800 nodes. Among all of 800 genes that I selected as highly differentially expressed genes, some could not be imported into the algorithm. In addition, the number of imported genes and terminal counts in a constructed networks are not equal as shown in **Table 3.1** and this represent the elimination of some terminals during optimization. Moreover, I performed a two way Mann-Whitney U test between terminals and intermediate nodes' degrees (**Figure 3.8**).

Table 3.1: Characteristics of networks. "Imported genes" is the number of differentially expressed genes (among 800) that are represented in the PPI dataset. "Network size" is the number of nodes in the optimal network. "Terminal count" represents the number of terminals included in the final network. The range i have tested for β were between 5 to 100 and ω range were between 1 to 10.

Datasets	**Imported Genes**	**Network Size**	**Terminal Counts**	**Intermediate Nodes**	β	ω
Berchtold_HIP	570	750	570	180	100	9
Berchtold$_SFG$	611	815	607	208	15	9
Berchtold_PCG	579	794	578	216	100	9
Lu_FC	711	860	710	150	100	9
Jonker_Brain	633	833	631	202	10	10
Narayanan_PFC	650	852	649	203	10	10
Zhang_CR	661	858	659	199	15	8
Zhang_VC	668	873	667	206	100	10
Zhang_PFC	658	868	657	211	10	10

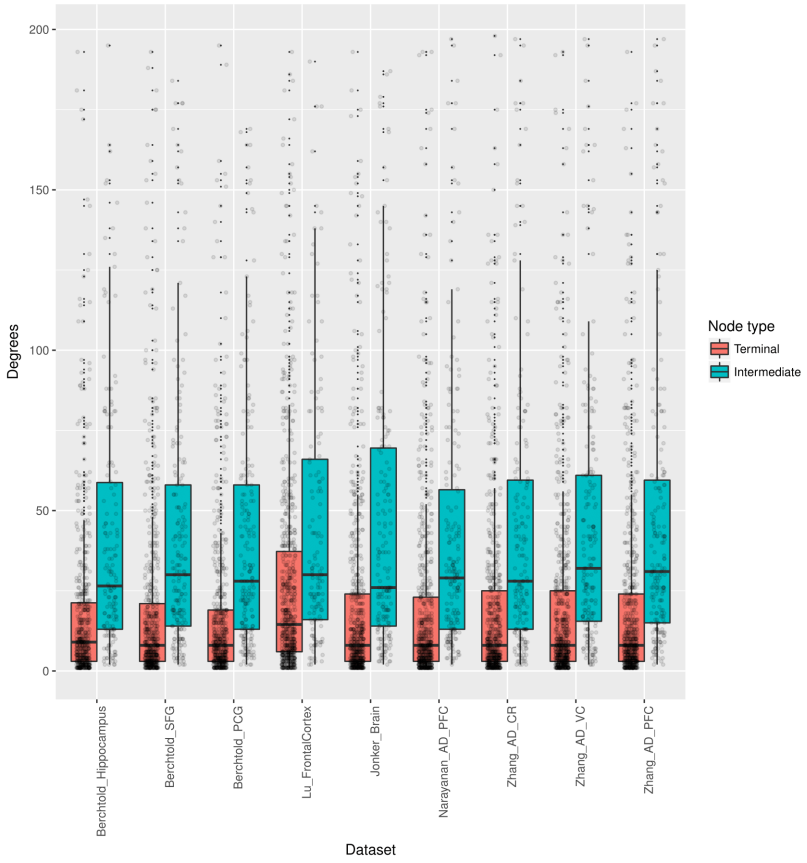

Figure 3.8: Degree distribution of intermediate and terminals nodes in each dataset. Orange boxplots represent terminals and green boxplots represent intermediate nodes. The Y axis is limited to 200. All comparisons are significant at MWU test p<0.001.

Degree distribution of terminal nodes are significantly lower than those of intermediate nodes. It is also important to state that there are many one degree nodes among terminals.

3.4 Clustering

I am assuming that, biological related genes are clustered together in the PPI network. In order to find communities in each network, I performed louvain modularity detection. This method initially calculates the modularity from two nodes and extends the connection to neighbor in order to find maximum modularity (**Figure 3.9**). I summarize the clustering outputs in **Table 3.2**.

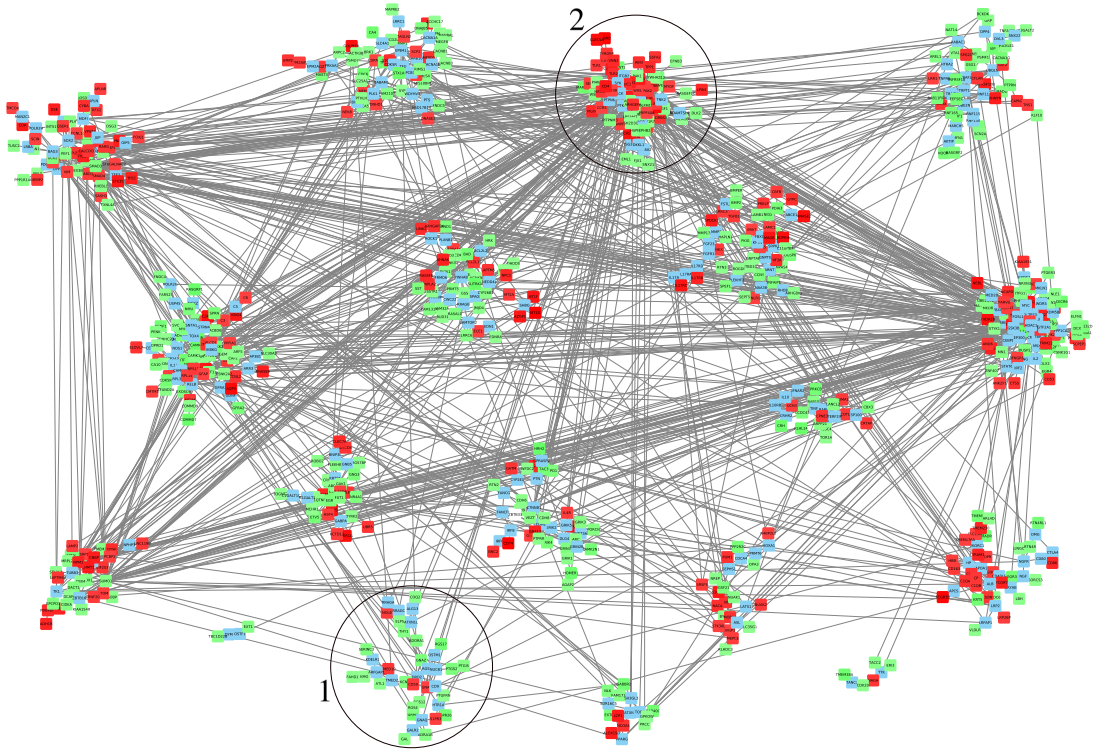

Figure 3.9: Clustering of Berchtold_PCG dataset. Louvain community detection algorithm calculate the local density of connected nodes within community compare to their connection in random network. Each separate group represent a component. Blue, red and green nodes represent intermediate nodes, up-regulated and down-regulated genes respectively. Circles are selected components which their functional enrichments will be explained in next section.

Table 3.2: Clustering of datasets. “Clusters count” represents the number of communities louvain modularity detected. “Removed clusters” is the number of clusters contain below 20 number of genes. “Min size” and “Max size” give the size of smallest and biggest clusters respectively. “sd size” represents the standard deviation of cluster sizes.

Datasets	**Clusters Count**	**Removed Clusters**	**Min Size**	**Max Size**	**std Size**
Berchtold_HIP	15	2	26	103	23.1
Berchtold$_SFG$	18	1	26	95	22.3
Berchtold_PCG	18	3	24	85	19.7
Lu_FC	15	1	20	115	26.5
Jonker_Brain	15	0	22	105	26.3
Narayanan_PFC	18	3	20	99	24.4
Zhang_CR	16	2	22	119	29.5
Zhang_VC	17	3	25	101	24
Zhang_PFC	16	0	21	103	25

3.5 Functional enrichment analysis

I performed enrichment analysis of KEGG pathway and Gene Ontology of clusters as I explained in **section 2.10**. To illustrate, the pathway which enriched in KEGG pathway enrichment analysis of circle 1 in **Figure 3.9** is "Serotonergic synapse". In addition, "Natural killer cell mediated cytotoxicity", "Regulation of actin cytoskeleton", "Fc gamma R-mediated phagocytosis", "Proteoglycans in cancer", "Chemokine signaling pathway", "Focal adhesion", "Pathogenic Escherichia coli infection", "Endocytosis", "Adherens junction", "T cell receptor signaling pathway", "Bacterial invasion of epithelial cells", "PI3K-Akt signaling pathway", "Leishmaniasis", "Tuberculosis", "Legionellosis", "Shigellosis" and "Salmonella infection" are pathways which enriched in circle 2 in **Figure 3.9**. KEGG pathway and Gene Ontology Biological Process enrichment results are given in **Figure 3.10** and **Figure 3.11** respectively. However, due to high number of enriched functional groups, I only demonstrate ones that were shared more than 4 times among 9 datasets and have a q-value lower than 0.1 for kegg and below 0.001 for Gene Ontology enrichment analysis. Functional groups which share separately among all AD or age networks are given in **Appendix A**.

Regulation of actin cytoskeleton is the only KEGG pathway which enriched in all datasets. Actin cytoskeleton preserve and maintain cell structure and has effect on polarity of cell. In addition, some studies indicated the relation of this microfilament with endocytosis and intracellular trafficking (Samaj et al., 2004). Another role of these filaments are in cell division and cytokinesis. Furthermore, actin cytoskeleton contribute in cell movement with the help of myosin. The organization of actin filaments are regulated by highly conserved actin-binding proteins. Some studies show that, distribution in regulation of actin cytoskeleton, such as mutations on actin or actin binding proteins leads to various disease such as cancer, cardiomyopathies and neurodegenerative diseases (Condeelis et al., 2005). In addition, increased actin turnover shown increase cell life span (Gourlay and Ayscough, 2005; Lee and Dominguez, 2010).

Revigo summarization of gene ontologies seen more than 4 times among datasets are given in **Figure 3.12**. Although "exocytosis" and "cell surface receptor signaling pathway" not enriched in all datasets, these two shared in most of datasets. Neurotransmission process is the secretion of neurotransmitters to the neural cleft and binding of these chemicals to the receptor of postsynaptic neuron to stimulate or inhibit neuronal activity. Exocytosis process plays important role in the secretion of neurotransmitters to the cleft. In addition, cell surface receptors of postsynaptic neuron are important in the initiate of signal transduction. Therefore, I believe disruption of these two functional groups cause problems in neuronal communication. I note that the activity of exocytosis is regulated by actin cytoskeleton (Porat-Shliom et al., 2013). Therefore changes in actin regulation may cause disruption in exocytosis process.

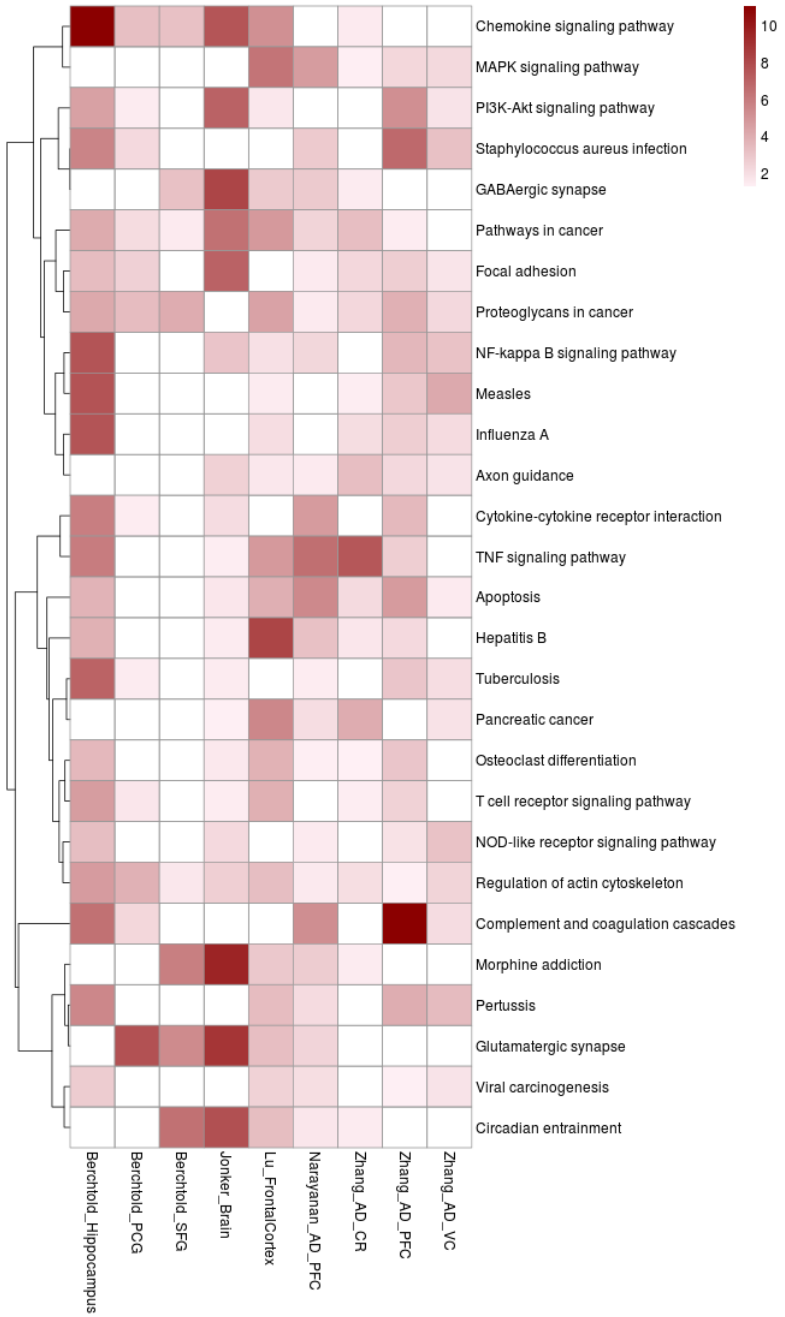

Figure 3.10: KEGG pathways enrichment analysis. The above heatmap only represents pathways seen more than 4 times among 9 datasets. Colors are log values of fischer test results. Dark red boxes represent highly significant *p*-values.

Figure 3.11: Gene Ontology Biological Process enrichment analysis. The above heatmap only represents pathways seen more than 4 times among 9 datasets. Dark red boxes represent highly significant *p*-values.

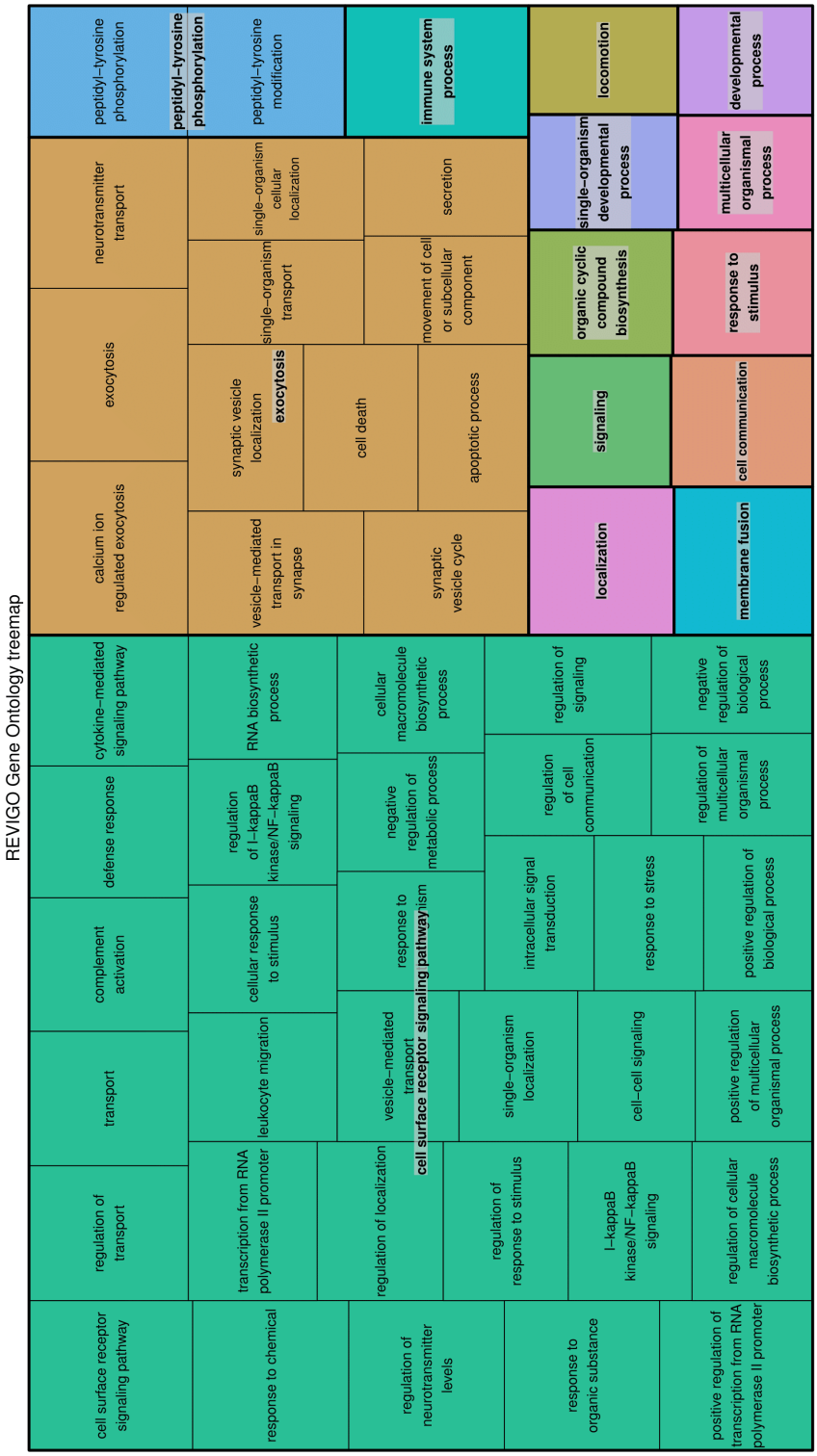

Figure 3.12: Revigo summarization of gene ontologies seen more than 4 times among datasets.

3.5.1 Permutation tests results

I can only evaluate the statistical and biological significance of functional groups by comparing with random permutation results. “Noisy Edges” is a permutation method which randomly adds noise to the edges. I believe that edge cost, terminal prizes and parameters shape the forest algorithm outcomes. Therefore, this “Noisy Edges” approach will change the results by altering edge costs. This helps us to clarify the probability of functional groups being enriched by chance. The “Aminoacyl-tRNA biosynthesis” in “Lu_FrontalCortex” is the only pathway that survived among all datasets compared to 100 permutation results. In addition, almost none of the Gene Ontologies survived among all datasets compare to 100 permutation results. Enrichment of same functional groups demonstrate the stability of network and hence indicate the robustness of edges to the noise.

“Shuffle Nodes” as its name implies, shuffles the prize of the nodes. Here, low prizes can represent high degree nodes. Therefore, due to the penalty for each connection, these genes can be eliminated. Thus, this could create a problem with the optimization of the forest and gives empty results. Therefore, in order to take also background genes into account I selected the lowest number of non-empty networks among datasets as the permutation count (n=29). **Figure 3.13** and **Figure 3.14** show KEGG pathway and Gene Ontology results which enriched significantly in compare of 29 “Shuffle Nodes” permutation.

Finally I performed a permutation of biological identifiers, such as age of individuals. In this permutation scheme, I shuffle ages (or AD status) in an expression matrix. This permutation helps us to reconstruct networks with genes which are classified as terminals simply by chance. Here again, I obtain some empty results for the parameters I inserted. Therefore, to apply enrichment analysis I took smallest number of non-empty networks among datasets as the a permutation count (n=12). **Figure 3.15** and **Figure 3.16** show KEGG pathway and Gene Ontology results for the age/AD permutation.

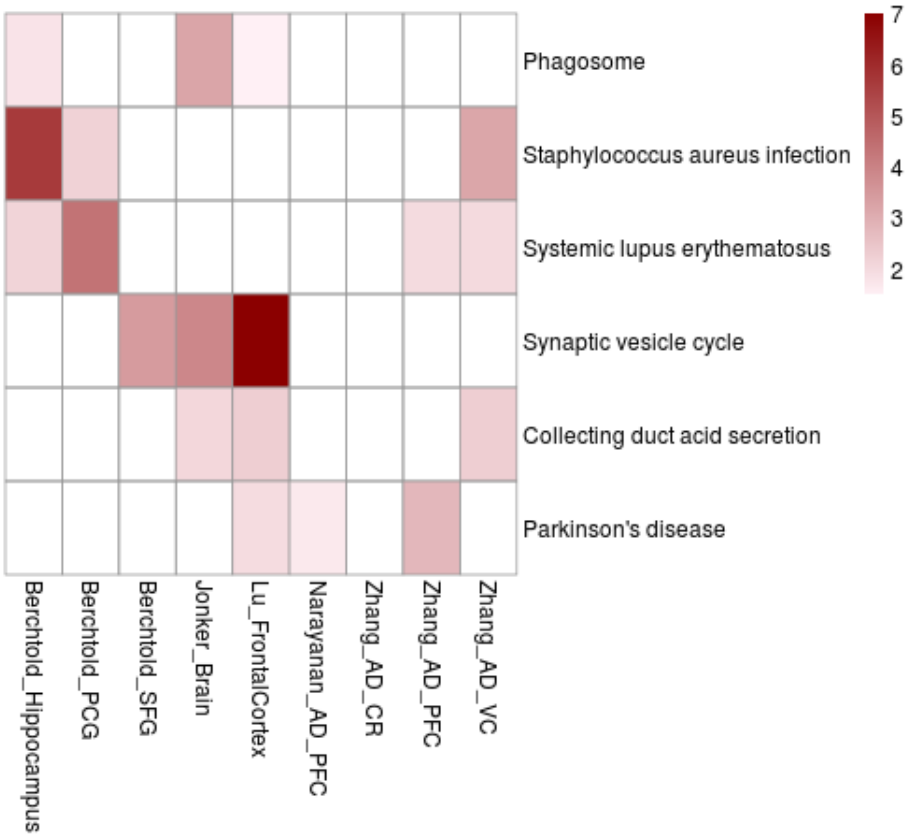

Figure 3.13: KEGG pathways significantly enriched in "Shuffle Prizes" permutation. The pathway which enriched more than 4 times among datasets is not exist. Therefore i exhibit pathways shown more than 2 times among datasets. Dark red color boxes represent highly significant *p*-values.

Figure 3.14: Gene Ontologies which significantly enriched compare to "Shuffle Prizes" Permutation. Gene Ontologies which enriched more than 4 times among datasets are not exist. Therefore i exhibit Gene Ontologies shown more than 2 times among datasets. Dark red color boxes represent highly significant p-values.

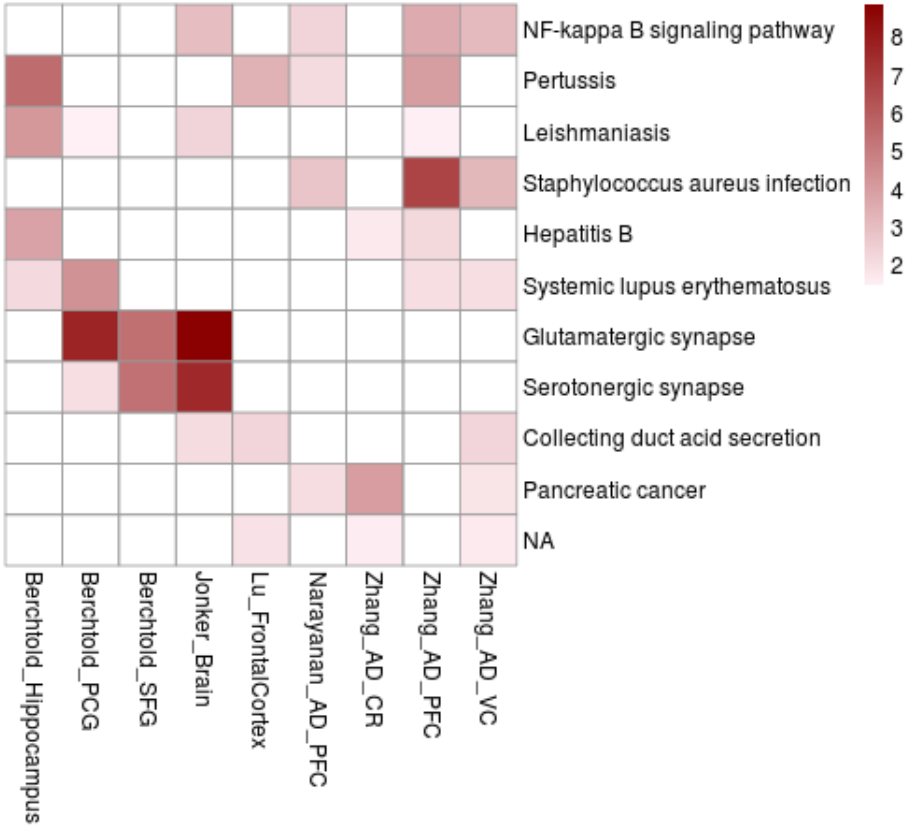

Figure 3.15: KEGG pathway which significantly enriched to 12 age/AD permutation results. The pathway which enriched more than 4 times among datasets is not exist. Therefore i exhibit pathways shown more than 2 times among datasets. Dark red color boxes represent highly significant *p*-values.

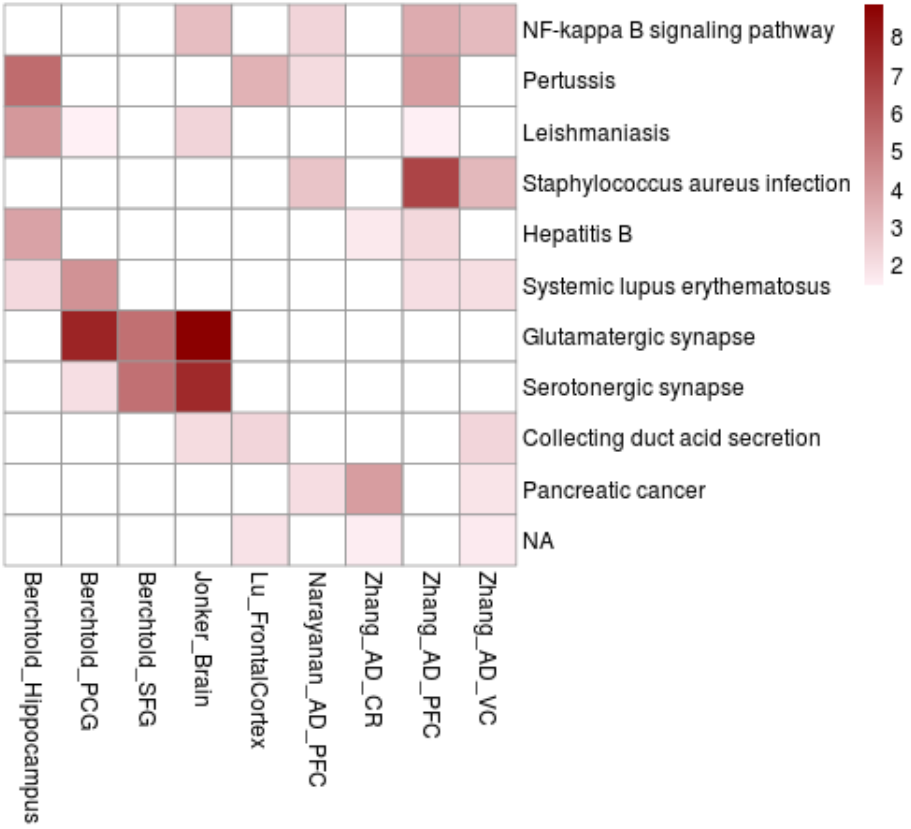

Figure 3.16: Gene Ontologies which significantly enriched compare to 12 age/AD permutation results. Gene Ontologies which enriched more than 4 times among datasets are not exist. Therefore i exhibit Gene Ontologies shown more than 2 times among datasets. Dark red color boxes represent highly significant *p*-values.

3.6 Common Edges

I tested whether there exist any interactions shared among all 9 datasets. I found no such case, and shared interactions were seen at most 7 times among all 9 datasets. Gene Ontology and KEGG pathway enrichment results of nodes supporting the idea that these interactions do not provide significant results (data not shown). This can be due to the low number of these nodes. The network in **Figure 3.17** represent interactions which represented more than 5 times among the 9 datasets. In addition genes which seen in all AD or age networks separately are given in **Appendix B**. Analyzing this network and investigate the characteristics of hubs are among the further studies of this project.

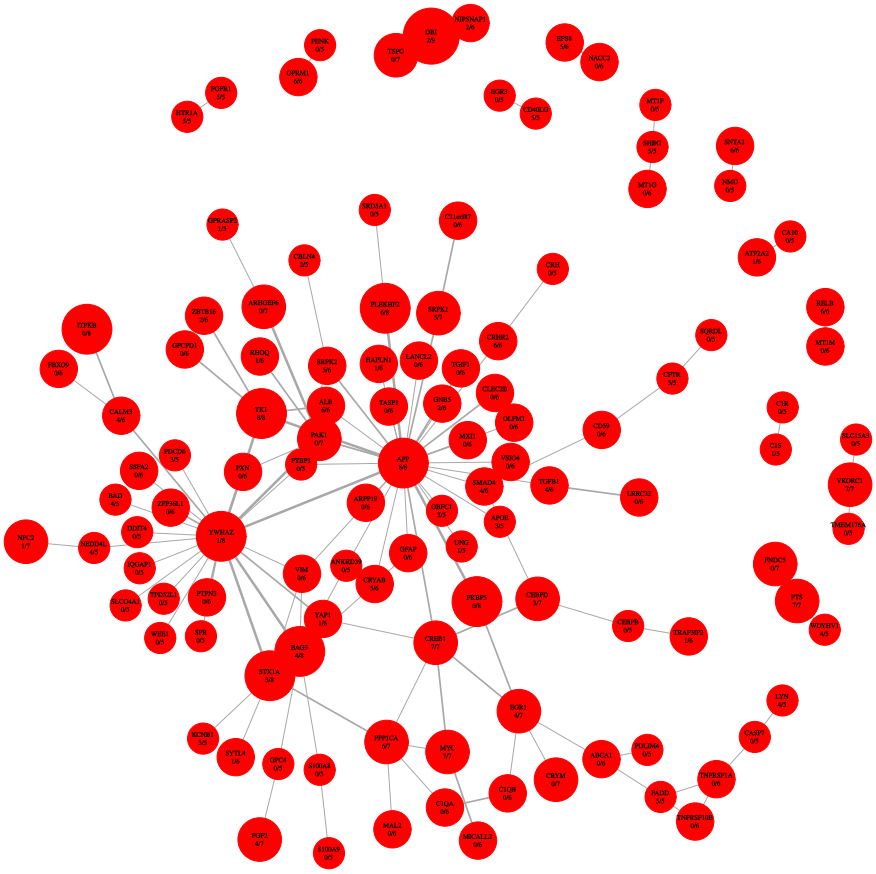

Figure 3.17: Common interactions among datasets. Above network is a connection between edges shared more than 5 times among datasets. Thickness of edges represent the amount of times this interactions seen in reconstructed networks and the size of nodes represent the number of reconstructed networks contains that gene. Labels are HGNC symbols and the fraction of intermediate nodes to the number of times that gene exist in reconstructed networks.

CHAPTER 4

DISCUSSION

Aging and age-related Alzheimer's disease are complex and heterogeneous processes. Therefore, in order to find genes which may participate in these two, just looking for gene expression changes in not sufficient. Network analysis will help us understand these genes' interactions in a biological network.

In this study, In order to construct the brain aging and Alzheimer's disease networks and search for common interactions they share in a network, I have used 5 aging and 4 AD microarray datasets as shown in **Tables 2.1** and **Table 2.2**. Subsequent to application of background corrections, normalization and multiple testing correction, I applied differential expression test using the Spearman correlation rank test. As it shown in **Figure 3.3**, AD datasets show high number of differentially expressed genes. Beside sample size, biological factors also may participate in the number of differential expression genes. Jonker_Brain contains the lowest amount of genes. In order to construct aging and AD networks I used the prize-collecting Steiner forest algorithm. The aim of this algorithm is to optimally connect selected genes within a protein-protein interaction network. It would be better to import all significant genes as input to the forest algorithm. However, the large input size could be a problem for this algorithm. Moreover, different amount of terminals in different datasets would rise bias in further analysis. To eliminate these problems, I anchored input size to 800.

To check consistency among datasets I applied pairwise Spearman correlation between them. As shown in **Figure 3.4**, AD and aging datasets cluster among them-

selves. It is also clear that, two PFC samples are highly correlated to each other. Except for the mouse brain dataset, Jonker_Brain, a slight correlation between aging and AD datasets can be seen. The PCA result in **Figure 3.5** clarifies the sharp separate clustering of AD and aging datasets. This result may be due to the common platforms they share. Therefore, it makes it difficult to distinguish biological and technical signal participate in this results. The Jonker_Brain dataset was identified as an outlier. Smear consistency of mice data with other datasets, stand off from other ones in a PCA and low amount of genes clarify the low quality of this dataset.

Before application of the forest algorithm, I checked for characteristics of the protein-protein interaction dataset. The degree distribution can clarify the biological properties of the database. The degree distribution was observed to follow a power law, and is thus classified as a scale-free network (Barabasi and Albert, 1999). This means that the number of nodes which are only connect to few neighbors is much higher than a genes connected to a high number of nodes. It is believed that there are few genes demonstrating hub characteristics in a biological networks. The structure shown clearly in **Figure 3.6** indicates that iRefIndex is a scale-free network. In a non−biological random network, degree distribution is normal (Costa et al., 2008).

After the application of prize-collecting Steiner forest algorithm using multiple combinations of parameters for each dataset I selected the largest network with the highest number of terminals. **Table 3.1** harbors these networks characteristics. This table explains that some of reconstructed networks are larger than 800 genes. This indicates the integration of intermediate nodes in a network. In addition, among all 800 terminal gene hits, some could not be imported to the algorithm. The reason is that some genes are not represented in the PPI database. Moreover, some genes may also be eliminated during optimization.

I also checked the degree distribution of terminal and Steiner nodes in a selected networks. In all datasets, the terminal nodes show significantly lower degree distribution than other nodes. I believe that the reason of this observation is the high amount of single degree terminal nodes. On the other hand, intermediate nodes have a role in connecting terminals, therefore edge nodes are always terminal ones. I should state

that one degree Steiner nodes would be eliminated from the network as it cause the increase of cost in objective function.

I cluster the selected network with the help of louvain modularity. The number of cluster and the maximum, minimum and standard deviation of cluster sizes are given in **Table 3.2**.

I performed GO and KEGG pathway enrichment analysis on the clusters for each datasets. **Figure 3.10** and **Figure 3.11** are heatmaps representing KEGG pathway and Gene Ontology Biological Process enrichment analysis results. Among multiple of KEGG pathways the only one shared in all datasets is “Regulation of actin cytoskeleton”. It is known that disruption in this regulation leads to various disease, such as neurodegenerative disease. Previous research determined that actin cytoskeleton has relation with intracellular signaling which regulates cellular activity and programmed cell death (Amberg et al., 2011). It is known that reactive oxygen species’ (ROS) accumulation leads to problems in mitochondria signaling and cell fate. Hence, actin cytoskeleton changes have been detected due to this accumulation (Gourlay and Ayscough, 2005). Revigo summarization of gene ontologies are given in **Figure 3.12**. I believe changes in "exocytosis" and "cell surface receptor signaling pathway" disrupt neuronal communication.

To check the significance of these enrichments, I applied three kinds of permutation, “Noisy Edges”, “Shuffle Prizes” and age permutation. Interestingly, all of the permutations reject the significant enrichments we observe in our results. The failure of “Noisy Edges” demonstrate the robustness of edges in a network. Increase of the noise could determine the level of robustness of these edges in a network. I am selecting most 800 significant differentially expressed genes. The variance of correlation coefficient values are low. Therefore, shuffle the prizes in “shuffle prizes” permutation does not change the network too much. In the age permutation, terminals change and the degree of these terminals also change in a network. This result, may claim that we could not interpret biological implication from this network. I was planning to look for common functional groups’ genes in a GBM dataset. However, as I am not able to detect common functional groups, I couldn’t use GBM data.

I construct the network in **Figure 3.17**, representing edges observed more than 5 times among datasets. In further studies, analysing this network's hubs could give us information about genes which play important role in aging and AD networks.

4.1 Limitations of the Study

In this study, unlike PFC, I used only one dataset for each brain region. Therefore, more than one dataset for each region and condition will help us to obtain more confidential results and allow to separate condition from tissue effect. Another limitation in this study is, I observed only one common functional group among all aging and AD datasets. Therefore, it is also important to analysing aging and AD datasets seperately to find conditional related functional groups. Randomization results obtain from "Shuffle prizes" demonstrate that some of trials were failed to reconstruct optimal network. Hence, the number of non-empty networks did not reach 100. Therefore, I should continuously permute the network until I achieve 100 non-empty ones. Another limitation is, inserting abs values to the algorithm, prevent us to analysis increased and decreased genes seperately. Therefore, it is better to examine them in an algorithm seperately.

CHAPTER 5

CONCLUSION

The aging phenotype is thought to involve expression changes in multiple genes, indicating the importance of studying interaction among genes rather than focusing on a single gene. I believe that, investigating common mechanisms in aging and AD in our network analysis method may help us eliminate possible inefficiency in microarray data, such as difficulties in detecting low light intensity signals, imperfect RNA hybridization or loss of gene expression information in meta-analysis comprising different platforms. Missing nodes due to technical effects can be readed in this approach. Another biological problem can be post-translational modifications of a protein which may affect expression of other genes and is not detected on a microarray. Although genes showing age-related expression changes and their interactions have been reported by several studies, these could also be studied using the Forest module. Using this method, we created an optimal interaction network of age-related or AD-associated genes in protein-protein interaction networks, by taking into consideration the terminal prizes and interaction strength of genes. These connections can be direct or be represented by intermediate nodes. In addition, we performed network alignment to test whether common interactions might be found in different species' and tissues' aging networks. The pathways common among all datasets was identified to be "regulation of actin cytoskeleton". However, Gene Ontology enrichment analysis did not provide shared functional groups. Further, to test the significance of the predicted interactions we used permutation, "Noisy Edges", "Shuffle Prizes" and age/AD permutation. Compared to these permutations most of the enrichments did not appear significant. This could be due to the insufficiency of permutations. I

believe that, the noise added to the edges demonstrated the robustness of edges in a network. On the other hand, terminal prizes are top 800 genes' correlation coefficients and these values are close to each other. Therefore, in "Shuffle Prizes" permutation, low amount of changes in terminal prizes does not alter network.

REFERENCES

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

Alzheimer's Association (2011). 2011 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 7(2):208–244.

Amberg, D., Leadsham, J. E., Kotiadis, V., and Gourlay, C. W. (2011). Cellular Ageing and the Actin Cytoskeleton. In *Sub-cellular biochemistry*, volume 57, pages 331–352.

Aoki, K. F. and Kanehisa, M. (2005). Using the KEGG Database Resource. In *Current Protocols in Bioinformatics*, volume Chapter 1, page Unit 1.12. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Armanios, M., Alder, J. K., Parry, E. M., Karim, B., Strong, M. A., and Greider, C. W. (2009). Short telomeres are sufficient to cause the degenerative defects associated with aging. *American journal of human genetics*, 85(6):823–32.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., Sherlock, G., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–9.

Avramopoulos, D., Szymanski, M., Wang, R., and Bassett, S. (2011). Gene expression reveals overlap between normal aging and Alzheimer's disease genes. *Neurobiology of Aging*, 32(12):2319.e27–2319.e34.

Baker, D. J., Wijshake, T., Tchkonia, T., LeBrasseur, N. K., Childs, B. G., van de Sluis, B., Kirkland, J. L., and van Deursen, J. M. (2011). Clearance of p16Ink4a-positive senescent cells delays ageing-associated disorders. *Nature*, 479(7372):232–236.

Barabasi and Albert (1999). Emergence of scaling in random networks. *Science (New York, N.Y.)*, 286(5439):509–12.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.

Blasco, M. A., Lee, H. W., Hande, M. P., Samper, E., Lansdorp, P. M., DePinho, R. A., and Greider, C. W. (1997). Telomere shortening and tumor formation by mouse cells lacking telomerase RNA. *Cell*, 91(1):25–34.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Brookmeyer, R., Johnson, E., Ziegler-Graham, K., and Arrighi, H. M. (2007). Forecasting the global burden of Alzheimer's disease. *Alzheimer's & Dementia*, 3(3):186–191.

Brunet, A. and Berger, S. L. (2014). Epigenetics of Aging and Aging-related Disease. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 69(Suppl 1):S17–S20.

Budovsky, A., Abramovich, A., Cohen, R., Chalifa-Caspi, V., and Fraifeld, V. (2007). Longevity network: Construction and implications. *Mechanisms of Ageing and Development*, 128(1):117–124.

Campisi, J. and d'Adda di Fagagna, F. (2007). Cellular senescence: when bad things happen to good cells. *Nature Reviews Molecular Cell Biology*, 8(9):729–740.

Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM*, 58(3):1–37.

Carvalho, B. S. and Irizarry, R. A. (2010). A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, 26(19):2363–2367.

Castilho, R. M., Squarize, C. H., Chodosh, L. A., Williams, B. O., and Gutkind, J. S. (2009). mTOR Mediates Wnt-Induced Epidermal Stem Cell Exhaustion and Aging. *Cell Stem Cell*, 5(3):279–289.

Chen, C., Liu, Y., Liu, Y., and Zheng, P. (2009). mTOR Regulation and Therapeutic Rejuvenation of Aging Hematopoietic Stem Cells. *Science Signaling*, 2(98):ra75–ra75.

Cokayne, K. (2003). *Experiencing old age in ancient Rome*. Routledge.

Condeelis, J., Singer, R. H., and Segall, J. E. (2005). THE GREAT ESCAPE: When Cancer Cells Hijack the Genes for Chemotaxis and Motility. *Annual Review of Cell and Developmental Biology*, 21(1):695–718.

Costa, L. d. F., Rodrigues, F. A., and Cristino, A. S. (2008). Complex networks: the key to systems biology. *Genetics and Molecular Biology*, 31(3):591–601.

Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., Bunney, W. E., Myers, R. M., Speed, T. P., Akil, H., Watson, S. J., and Meng, F. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Research*, 33(20):e175–e175.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210.

Ferrarini, L., Bertelli, L., Feala, J., McCulloch, A. D., and Paternostro, G. (2005). A more efficient search strategy for aging genes based on connectivity. *Bioinformatics*, 21(3):338–348.

Fontana, L., Partridge, L., and Longo, V. D. (2010). Extending Healthy Life Span–From Yeast to Humans. *Science*, 328(5976):321–326.

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3):307–315.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80.

Gerhold, D., Rushmore, T., and Caskey, C. (1999). DNA chips: promising toys have become powerful tools. *Trends in Biochemical Sciences*, 24(5):168–173.

Gitter, A. and Bar-Joseph, Z. (2013). Identifying proteins controlling key disease signaling pathways. *Bioinformatics*, 29(13):i227–i236.

Gosline, S. J. C., Spencer, S. J., Ursu, O., and Fraenkel, E. (2012). SAMNet: a network-based approach to integrate multi-dimensional high throughput datasets. *Integrative Biology*, 4(11):1415.

Gourlay, C. W. and Ayscough, K. R. (2005). Opinion: The actin cytoskeleton: a key regulator of apoptosis and ageing? *Nature Reviews Molecular Cell Biology*, 6(7):583–589.

Hauke, J. and Kossowski, T. (2011). Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. *Quaestiones Geographicae*, 30(2):87–93.

Hayflick, L. and Moorhead, P. (1961). The serial cultivation of human diploid cell strains. *Experimental Cell Research*, 25(3):585–621.

Herrera, E., Samper, E., Martín-Caballero, J., Flores, J. M., Lee, H. W., and Blasco, M. A. (1999). Disease states associated with telomerase deficiency appear earlier in mice with short telomeres. *The EMBO Journal*, 18(11):2950–2960.

Hicks, S. C. and Irizarry, R. A. (2014). When to use Quantile Normalization? *bioRxiv*.

Hsieh, S. (2015). The Importance of Aging Brain Research. *Journal of Gerontology & Geriatric Research*, 04(02).

Hubbell, E., Liu, W.-M., and Mei, R. (2002). Robust estimators for expression analysis. *Bioinformatics*, 18(12):1585–1592.

Irizarry, R. A., Hobbs, B., Collin, F., BeazerBarclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264.

Johnson, S. C., Dong, X., Vijg, J., and Suh, Y. (2015). Genetic evidence for common pathways in human age-related diseases. *Aging cell*, 14(5):809–17.

Jonker, M. J., Melis, J. P. M., Kuiper, R. V., van der Hoeven, T. V., Wackers, P. F. K., Robinson, J., van der Horst, G. T. J., Dollé, M. E. T., Vijg, J., Breit, T. M., Hoeijmakers, J. H. J., and van Steeg, H. (2013). Life spanning murine gene expression profiles in relation to chronological and pathological aging in multiple organs. *Aging Cell*, 12(5):901–909.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30.

Kanfi, Y., Naiman, S., Amir, G., Peshti, V., Zinman, G., Nahum, L., Bar-Joseph, Z., and Cohen, H. Y. (2012). The sirtuin SIRT6 regulates lifespan in male mice. *Nature*, 483(7388):218–221.

Klass, M. R. (1983). A method for the isolation of longevity mutants in the nematode Caenorhabditis elegans and initial results. *Mechanisms of ageing and development*, 22(3-4):279–86.

Koga, H., Kaushik, S., and Cuervo, A. M. (2011). Protein homeostasis and aging: The importance of exquisite quality control. *Ageing Research Reviews*, 10(2):205–215.

Lee, S. H. and Dominguez, R. (2010). Regulation of actin cytoskeleton dynamics in cells. *Molecules and cells*, 29(4):311–325.

Liu, H., Bebu, I., and Li, X. (2010). Microarray probes and probe sets. *Frontiers in bioscience (Elite edition)*, 2:325–38.

López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The hallmarks of aging. *Cell*, 153(6):1194–217.

Matheu, A., Maraver, A., Collado, M., Garcia-Cao, I., Cañamero, M., Borras, C., Flores, J. M., Klatt, P., Viña, J., and Serrano, M. (2009). Anti-aging activity of the Ink4/Arf locus. *Aging Cell*, 8(2):152–161.

Matheu, A., Maraver, A., Klatt, P., Flores, I., Garcia-Cao, I., Borras, C., Flores, J. M., Viña, J., Blasco, M. A., and Serrano, M. (2007). Delayed ageing through damage protection by the Arf/p53 pathway. *Nature*, 448(7151):375–379.

Mazin, P., Xiong, J., Liu, X., Yan, Z., Zhang, X., Li, M., He, L., Somel, M., Yuan, Y., Phoebe Chen, Y.-P., Li, N., Hu, Y., Fu, N., Ning, Z., Zeng, R., Yang, H., Chen, W., Gelfand, M., and Khaitovich, P. (2013). Widespread splicing changes in human brain development and aging. *Molecular Systems Biology*, 9(1):633–633.

Miller, J. A., Oldham, M. C., and Geschwind, D. H. (2008). A Systems Level Analysis of Transcriptional Changes in Alzheimer's Disease and Normal Aging. *Journal of Neuroscience*, 28(6):1410–1420.

Moskalev, A. A., Shaposhnikov, M. V., Plyusnina, E. N., Zhavoronkov, A., Budovsky, A., Yanai, H., and Fraifeld, V. E. (2013). The role of DNA damage and repair in aging through the prism of Koch-like criteria. *Ageing Research Reviews*, 12(2):661–684.

Mostoslavsky, R., Chua, K. F., Lombard, D. B., Pang, W. W., Fischer, M. R., Gellon, L., Liu, P., Mostoslavsky, G., Franco, S., Murphy, M. M., Mills, K. D., Patel, P., Hsu, J. T., Hong, A. L., Ford, E., Cheng, H.-L., Kennedy, C., Nunez, N., Bronson, R., Frendewey, D., Auerbach, W., Valenzuela, D., Karow, M., Hottiger, M. O., Hursting, S., Barrett, J. C., Guarente, L., Mulligan, R., Demple, B., Yancopoulos,

G. D., and Alt, F. W. (2006). Genomic Instability and Aging-like Phenotype in the Absence of Mammalian SIRT6. *Cell*, 124(2):315–329.

Olovnikov, A. M. (1996). Telomeres, telomerase, and aging: origin of the theory. *Experimental gerontology*, 31(4):443–8.

Pal, S. and Tyler, J. K. (2016). Epigenetics and aging. *Science Advances*, 2(7).

Park, C. B. and Larsson, N.-G. (2011). Mitochondrial DNA mutations in disease and aging. *The Journal of cell biology*, 193(5):809–18.

Parmigiani, G., Garrett, E. S., Irizarry, R. A., and Zeger, S. L. (2003). The Analysis of Gene Expression Data: An Overview of Methods and Software. pages 1–45. Springer, New York, NY.

Patil, A. and Nakai, K. (2014). TimeXNet: Identifying active gene sub-networks using time-course gene expression profiles. *BMC Systems Biology*, 8(Suppl 4):S2.

Paull, E. O., Carlin, D. E., Niepel, M., Sorger, P. K., Haussler, D., and Stuart, J. M. (2013). Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics*, 29(21):2757–2764.

Porat-Shliom, N., Milberg, O., Masedunskas, A., and Weigert, R. (2013). Multiple roles for the actin cytoskeleton during regulated exocytosis. *Cellular and molecular life sciences : CMLS*, 70(12):2099–121.

Promislow, D. E. L. (2004). Protein networks, pleiotropy and the evolution of senescence. *Proceedings of the Royal Society B: Biological Sciences*, 271(1545):1225–1234.

Rera, M., Bahadorani, S., Cho, J., Koehler, C. L., Ulgherait, M., Hur, J. H., Ansari, W. S., Lo, T., Jones, D. L., Walker, D. W., and Walker, D. W. (2011). Modulation of longevity and tissue homeostasis by the Drosophila PGC-1 homolog. *Cell metabolism*, 14(5):623–34.

Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology*, 26(3):303–304.

Rousseeuw, P. J., Ruts, I., and Tukey, J. W. (1999). The Bagplot: A Bivariate Boxplot. *The American Statistician*, 53(4):382–387.

Rudolph, K. L., Chang, S., Lee, H. W., Blasco, M., Gottlieb, G. J., Greider, C., and DePinho, R. A. (1999). Longevity, stress response, and cancer in aging telomerase-deficient mice. *Cell*, 96(5):701–12.

Salminen, A., Kaarniranta, K., and Kauppinen, A. (2012). Inflammaging: disturbed interplay between autophagy and inflammasomes. *Aging*, 4(3):166–175.

Samaj, J., Baluska, F., Voigt, B., Schlicht, M., Volkmann, D., and Menzel, D. (2004). Endocytosis, actin cytoskeleton, and signaling. *Plant physiology*, 135(3):1150–61.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray.

Tissenbaum, H. A. (2015). Using C. elegans for aging research. *Invertebrate reproduction & development*, 59(sup1):59–63.

Tomás-Loba, A., Flores, I., Fernández-Marcos, P. J., Cayuela, M. L., Maraver, A., Tejera, A., Borrás, C., Matheu, A., Klatt, P., Flores, J. M., Viña, J., Serrano, M., and Blasco, M. A. (2008). Telomerase Reverse Transcriptase Delays Aging in Cancer-Resistant Mice. *Cell*, 135(4):609–622.

Tuncbag, N., Braunstein, A., Pagnani, A., Huang, S.-S. C., Chayes, J., Borgs, C., Zecchina, R., and Fraenkel, E. (2013). Simultaneous Reconstruction of Multiple Signaling Pathways via the Prize-Collecting Steiner Forest Problem. *Journal of Computational Biology*, 20(2):124–136.

Tuncbag, N., Gosline, S. J. C., Kedaigle, A., Soltis, A. R., Gitter, A., and Fraenkel, E. (2016). Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. *PLOS Computational Biology*, 12(4):e1004879.

Turner, B., Razick, S., Turinsky, A. L., Vlasblom, J., Crowdy, E. K., Cho, E., Morrison, K., Donaldson, I. M., and Wodak, S. J. (2010). iRefWeb: interactive analysis

of consolidated protein interaction data and their supporting evidence. *Database*, 2010(0):baq023–baq023.

Vandin, F., Upfal, E., and Raphael, B. J. (2011). Algorithms for Detecting Significantly Mutated Pathways in Cancer. *Journal of Computational Biology*, 18(3):507–522.

Verhoeven, K. J., Simonsen, K. L., and McIntyre, L. M. (2005). Implementing false discovery rate control: increasing your power. *Oikos*, 108(3):643–647.

Vijg, J. and Suh, Y. (2013). Genome Instability and Aging. *Annual Review of Physiology*, 75(1):645–668.

Villaveces, J. M., Jimenez, R. C., Porras, P., Del-Toro, N., Duesbury, M., Dumousseau, M., Orchard, S., Choi, H., Ping, P., Zong, N. C., Askenazi, M., Habermann, B. H., and Hermjakob, H. (2015). Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study. *Database*, 2015(0):bau131–bau131.

Wang, J., Zhang, S., Wang, Y., Chen, L., and Zhang, X.-S. (2009). Disease-Aging Network Reveals Significant Roles of Aging Genes in Connecting Genetic Diseases. *PLoS Computational Biology*, 5(9):e1000521.

Whitlock, M. and Schluter, D. (2009). *The analysis of biological data*. Roberts and Co. Publishers.

Yeger-Lotem, E., Riva, L., Su, L. J., Gitler, A. D., Cashikar, A. G., King, O. D., Auluck, P. K., Geddie, M. L., Valastyan, J. S., Karger, D. R., Lindquist, S., and Fraenkel, E. (2009). Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nature genetics*, 41(3):316–23.

Yilmaz, Ö. H., Katajisto, P., Lamming, D. W., Gültekin, Y., Bauer-Rowe, K. E., Sengupta, S., Birsoy, K., Dursun, A., Yilmaz, V. O., Selig, M., Nielsen, G. P., Mino-Kenudson, M., Zukerberg, L. R., Bhan, A. K., Deshpande, V., and Sabatini, D. M. (2012). mTORC1 in the Paneth cell niche couples intestinal stem-cell function to calorie intake. *Nature*, 486(7404):490–5.

Yuan, Y., Chen, Y. P. P., Boyd-Kirkup, J., Khaitovich, P., and Somel, M. (2012). Accelerated aging-related transcriptome changes in the female prefrontal cortex. *Aging Cell*, 11(5):894–901.

APPENDIX A

LIST OF SHARED FUNCTIONAL GROUPS AMONG AD AND AGING NETWORKS SEPARATELY

Table A.1: List of KEGG pathways shared among aging datasets.

KEGG ID	Name
hsa04062	Chemokine signaling pathway
hsa04810	Regulation of actin cytoskeleton
hsa05200	Pathways in cancer

Table A.2: List of KEGG pathways shared among AD datasets.

KEGG ID	Name
hsa04210	Apoptosis
hsa04217	Necroptosis
hsa04510	Focal adhesion
hsa04810	Regulation of actin cytoskeleton
hsa05205	Proteoglycans in cancer
hsa04360	Axon guidance
hsa04010	MAPK signaling pathway

Table A.3: List of GO Biological Process categories shared among aging datasets.

GO ID	GO Term

GO:0007154	cell communication
GO:0007165	signal transduction
GO:0023052	signaling
GO:0044700	single organism signaling
GO:0044765	single-organism transport
GO:0050896	response to stimulus
GO:0051716	cellular response to stimulus
GO:1902578	single-organism localization

Table A.4: List of GO Biological Process categories shared among AD datasets.

GO ID	GO Term
GO:0006915	apoptotic process
GO:0008219	cell death
GO:0008625	extrinsic apoptotic signaling pathway via death domain receptors
GO:0010558	negative regulation of macromolecule biosynthetic process
GO:0012501	programmed cell death
GO:0034097	response to cytokine
GO:0051253	negative regulation of RNA metabolic process
GO:1902679	negative regulation of RNA biosynthetic process
GO:1903507	negative regulation of nucleic acid-templated transcription

APPENDIX B

LIST OF SHARED GENES AMONG AD AND AGING NETWORKS SEPARATELY

Table B.1: List of genes shared among all aging datasets.

ENSG ID	Gene Name
ENSG00000172531	PPP1CA
ENSG00000155368	DBI

Table B.2: List of genes shared among human aging datasets.

ENSG ID	Gene Name
ENSG00000088826	SMOX
ENSG00000096060	FKBP5
ENSG00000112425	EPM2A
ENSG00000112559	MDFI
ENSG00000124942	AHNAK
ENSG00000125144	MT1G
ENSG00000129214	SHBG
ENSG00000143772	ITPKB
ENSG00000155368	DBI
ENSG00000162728	KCNJ9
ENSG00000164924	YWHAZ
ENSG00000172531	PPP1CA
ENSG00000175895	PLEKHF2
ENSG00000178567	EPM2AIP1

ENSG00000183763	TRAIP
ENSG00000196616	ADH1B
ENSG00000198417	MT1F
ENSG00000213853	EMP2

Table B.3: List of genes shared among all AD datasets.

ENSG ID	**Gene Name**
ENSG00000001626	CFTR
ENSG00000005022	SLC25A5
ENSG00000011304	PTBP1
ENSG00000054803	CBLN4
ENSG00000056736	IL17RB
ENSG00000056972	TRAF3IP2
ENSG00000065882	TBC1D1
ENSG00000067182	TNFRSF1A
ENSG00000070159	PTPN3
ENSG00000070831	CDC42
ENSG00000073536	NLE1
ENSG00000075415	SLC25A3
ENSG00000076716	GPC4
ENSG00000078043	PIAS2
ENSG00000085063	CD59
ENSG00000088812	ATRN
ENSG00000089123	TASP1
ENSG00000096060	FKBP5
ENSG00000100906	NFKBIA
ENSG00000101187	SLCO4A1
ENSG00000101276	SLC52A3
ENSG00000102362	SYTL4
ENSG00000103316	CRYM
ENSG00000103591	AAGAB
ENSG00000105329	TGFB1

ENSG00000106089	STX1A
ENSG00000106484	MEST
ENSG00000107872	FBXL15
ENSG00000108349	CASC3
ENSG00000109906	ZBTB16
ENSG00000110148	CCKBR
ENSG00000110852	CLEC2B
ENSG00000111652	COPS7A
ENSG00000111907	TPD52L1
ENSG00000112146	FBXO9
ENSG00000112739	PRPF4B
ENSG00000112818	MEP1A
ENSG00000113558	SKP1
ENSG00000113916	BCL6
ENSG00000115825	PRKD3
ENSG00000116044	NFE2L2
ENSG00000116353	MECR
ENSG00000117118	SDHB
ENSG00000118260	CREB1
ENSG00000118473	SGIP1
ENSG00000119655	NPC2
ENSG00000119950	MXI1
ENSG00000120063	GNA13
ENSG00000120875	DUSP4
ENSG00000120889	TNFRSF10B
ENSG00000121774	KHDRBS1
ENSG00000121858	TNFSF10
ENSG00000122584	NXPH1
ENSG00000124440	HIF3A
ENSG00000125148	MT2A
ENSG00000125772	GPCPD1
ENSG00000129116	PALLD
ENSG00000129473	BCL2L2
ENSG00000129675	ARHGEF6
ENSG00000130024	PHF10

ENSG00000130254	SAFB2
ENSG00000130770	ATPIF1
ENSG00000132329	RAMP1
ENSG00000132357	CARD6
ENSG00000132434	LANCL2
ENSG00000134569	LRP4
ENSG00000134575	ACP2
ENSG00000135250	SRPK2
ENSG00000136827	TOR1A
ENSG00000137210	TMEM14B
ENSG00000137507	LRRC32
ENSG00000137693	YAP1
ENSG00000137767	SQRDL
ENSG00000138411	HECW2
ENSG00000138685	FGF2
ENSG00000139910	NOVA1
ENSG00000139921	TMX1
ENSG00000141646	SMAD4
ENSG00000142192	APP
ENSG00000142227	EMP3
ENSG00000143727	ACP1
ENSG00000143772	ITPKB
ENSG00000145358	DDIT4L
ENSG00000146416	AIG1
ENSG00000146701	MDH2
ENSG00000147437	GNRH1
ENSG00000148411	NACC2
ENSG00000149269	PAK1
ENSG00000149573	MPZL2
ENSG00000150787	PTS
ENSG00000151491	EPS8
ENSG00000151929	BAG3
ENSG00000153914	SREK1
ENSG00000154582	TCEB1
ENSG00000155368	DBI

ENSG00000155659	VSIG4
ENSG00000156097	GPR61
ENSG00000158301	GPRASP2
ENSG00000159403	C1R
ENSG00000160097	FNDC5
ENSG00000162188	GNG3
ENSG00000162413	KLHL21
ENSG00000163032	VSNL1
ENSG00000163743	RCHY1
ENSG00000163884	KLF15
ENSG00000164332	UBLCP1
ENSG00000164761	TNFRSF11B
ENSG00000164924	YWHAZ
ENSG00000164949	GEM
ENSG00000165029	ABCA1
ENSG00000165704	HPRT1
ENSG00000165806	CASP7
ENSG00000166111	SVOP
ENSG00000166483	WEE1
ENSG00000167785	ZNF558
ENSG00000167900	TK1
ENSG00000168003	SLC3A2
ENSG00000168040	FADD
ENSG00000168874	ATOH8
ENSG00000169217	CD2BP2
ENSG00000169271	HSPB3
ENSG00000170035	UBE2E3
ENSG00000170370	EMX2
ENSG00000170899	GSTA4
ENSG00000171450	CDK5R2
ENSG00000172216	CEBPB
ENSG00000173039	RELA
ENSG00000173530	TNFRSF10D
ENSG00000175287	PHYHD1
ENSG00000175352	NRIP3

ENSG00000175895	PLEKHF2
ENSG00000176046	NUPR1
ENSG00000177426	TGIF1
ENSG00000177432	NAP1L5
ENSG00000177575	CD163
ENSG00000178252	WDR6
ENSG00000179915	NRXN1
ENSG00000181826	RELL1
ENSG00000182326	C1S
ENSG00000183943	PRKX
ENSG00000184117	NIPSNAP1
ENSG00000185022	MAFF
ENSG00000185519	FAM131C
ENSG00000185650	ZFP36L1
ENSG00000186951	PPARA
ENSG00000187193	MT1X
ENSG00000196954	CASP4
ENSG00000198604	BAZ1A
ENSG00000198890	PRMT6
ENSG00000198932	GPRASP1
ENSG00000206535	LNP1
ENSG00000213337	ANKRD39
ENSG00000213920	MDP1
ENSG00000214050	FBXO16
ENSG00000221869	CEBPD
ENSG00000239672	NME1
ENSG00000243364	EFNA4
ENSG00000254087	LYN