Susanne Glenna

A comparative study of gene correlation patterns and mean expression levels in Alzheimer's disease by network analysis

Master's thesis in Biotechnology

Supervisor: Almaas, Eivind

December 2020

Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

NTNU
Norwegian University of
Science and Technology

Susanne Glenna

A comparative study of gene correlation patterns and mean expression levels in Alzheimer's disease by network analysis

Master's thesis in Biotechnology
Supervisor: Almaas, Eivind
December 2020

Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

NTNU
Norwegian University of
Science and Technology

Summary

Every three seconds, someone in the world develops dementia [1]. Their brain cells die, memory is gradually lost, and familiar places become unfamiliar. It is devastating not only for the individual, but also for family and caretakers. This study sought to bring new insight into the pathology of Alzheimer's disease (AD), the most common cause of dementia.

Microarray-based technologies are widely used to study patterns of gene expression on a genome-wide scale [2]. The development of high-throughput experimental techniques along with the growth in computational power has enabled the study of thousands of genes in one experiment. It is now possible to investigate the interplay of genes responsible for complex diseases, such as AD, by analyzing the changes in gene co-expression patterns between healthy and disease states. In this work, the newly developed CSD (Conserved, Specific, Differentiated co-expression) method [3] was used on AD microarray data for this purpose.

The method successfully generated a differential co-expression network enriched in genes with AD-related functions. As much as 64 genes in the network were previously associated with AD, including one of the largest hubs; *VSNL1*. 17 other network hubs were identified: KIAA1841, NMNAT2, MIGA2, AQR, AL158206.1, HPRT1, GTF2I, TOM1L2, YWHAH, GOT1, NAPB, TMEM178A, PLTP, LCAT, ENPP2, CADPS and MDH1. Their high connectivity in the network and involvement in processes that are important for AD progress make them prominent genes for further studies. The most highly enriched pathways in the network were major signaling pathways essential for synaptic transmission, which when aberrant can ultimately lead to synaptic loss and cell death, key features of AD [4, 5].

Differential expression analysis (DEA) was added to the framework to look for changes in the regulations of *individual* genes as well. In combination with the co-expression analysis, DEA offered new insights both in terms of method development and for increased biological insight into AD. Most genes in the network were not differentially expressed genes (DEGs), which confirmed that a change in co-expression is not necessarily due to changes in mean expression of the correlated genes. Interestingly, the integrated analysis also revealed that the conserved (C)-type of co-expression is a valuable part of the CSD method and can even be directly interesting from a disease perspective. In particular, *levels* of gene expression were affected by the disease, mostly down-regulated, even if the pairwise *correlations* were conserved. This is novel compared to what has been assumed earlier when applying this method. More research is needed to validate these new findings, and to explore the underlying mechanisms behind the proposed disease gene candidates. In the end, the hope is that the identification of dysregulations at the transcriptome level will aid in the clinical diagnosis and treatment of AD.

Sammendrag

Hvert tredje sekund blir én person i verden rammet av demens [1]. Hjernecellene deres dør, hukommelsen går gradvis tapt og kjente steder blir ukjente. Det er ikke bare ødeleggende for individet selv, men påvirker også familien og andre omsorgspersoner. Dette studiet hadde til hensikt å gi ny innsikt i patologien til Alzheimers sykdom (AD), den vanligste årsaken til demens.

DNA-mikromatriser er mye brukt for å studere genuttrykksmøntre på genomskala [2]. Utviklingen av eksperimentelle teknikker med høy gjennomstrømning sammen med en stor vekst i datakraft har gjort det mulig å studere tusenvis av gener samtidig. Det er nå mulig å undersøke samspillet mellom gener som er ansvarlige for komplekse sykdommer, som AD, ved å analysere endringene i koekspresjonsmønstre fra frisk til syk. I dette arbeidet ble den nyutviklede CSD-metoden (konservert, spesifikk, differensiert koekspresjon) [3] brukt på AD-mikromatrisedata til dette formålet.

Metoden lyktes med å generere et differensielt koekspresjonsnettverk beriket med gener med AD-relaterte funksjoner. Så mye som 64 gener i nettverket var tidligere assosiert med AD, inkludert en av de største navene; *VSNL1*. 17 andre nettverksnav ble identifisert: KIAA1841, NMNAT2, MIGA2, AQR, AL158206.1, HPRT1, GTF2I, TOM1L2, YWHAH, GOT1, NAPB, TMEM178A, PLTP, LCAT, ENPP2, CADPS og MDH1. Deres kobling til mange gener i nettverket, samt involvering i prosesser relatert til sykdomsforløpet til AD, gjør dem til fremtredende kandidater for videre studier. De mest overrepresenterte reaksjonssporene i nettverket var involvert i overføring av nervesignaler, som når forstyrret kan ende i synapsetap og celledød, kritiske aspekter ved AD [4, 5].

Analyse av differensielt genuttrykk ble lagt til i rammeverket for å se etter endringer i regulering av *individuelle* gener i tillegg til korrelasjon mellom gen*par*. Dette ga ny innsikt, både med tanke på metodeutvikling og for økt biologisk innsikt i sykdommen. Flesteparten av genene i nettverket var ikke differensielt uttrykte gener (DEGer), som bekreftet at en endring i koekspresjon ikke nødvendigvis skyldes endringen i gjennomsnittlig uttrykk for de to korrelerte genene. I tillegg avslørte den integrerte analysen at den konserverte (C)-typen av koekspresjon er en verdifull del av CSD-metoden, som kan være direkte interessant fra et sykdomsperspektiv. Spesielt ble *nivåer* av genuttrykk påvirket av sykdommen, for det meste nedregulert, selv om de parvise *korrelasjonene* var bevarte. Dette er nytt sammenlignet med hva som er tidligere antatt ved bruk av denne metoden. Mer forskning er nødvendig for å validere disse nye funnene, samt for å utforske underliggende mekanismer bak de foreslåtte sykdomsforbindelsene. Håpet er at identifiseringen av dysreguleringer på transkriptomnivå til slutt kan bidra til klinisk diagnose og behandling av AD.

Preface

This thesis concludes my Master of Science degree in Biotechnology, with specialization in *Systems Biology*, at the Norwegian University of Science and Technology (NTNU) in Trondheim. The ways complex biological systems can be studied by simple, universal network parameters has left me astonished.

The year of 2020 has been challenging for everyone due to the raging COVID-19 pandemic. It has restricted us in several ways, but it has also enabled us to discover new ways of working together. Although it was favorable not to have laboratory work in this thesis, the pandemic still affected me personally. Thankfully, I was not alone in all of this.

First, I want to express my gratitude to all the brilliant people within the *Network Systems Biology* group, most of all to my supervisor Professor Eivind Almaas. His ability to see things in the bigger picture is what led me to the field of systems biology in the first place. I am grateful for all his advice and encouragement throughout this project. I would also like to thank my co-supervisor, Post.Doc. André Voigt, for his guidance in programming and statistics, especially in the implementation of the CSD method.

I also want to thank all my friends and classmates for all the fun times we have had over the years, and for supporting me through my challenges and meltdowns, especially during this last year. Specifically, I would like to thank Halvor Kvernes Meen for always being available to share his expertise in Python programming with me. I also want to explicitly thank Camilla Mauroy and Ada Nilsen Nordeidet for their guidance with this thesis and our scientific conversations over a glass of wine. Lastly, to all of you who believed in me when I did not believe in myself; I am forever grateful.

Susanne Glenna
Trondheim, December 2020

Table of Contents

List of Tables

List of Figures

Abbreviations

Aβ	=	Amyloid Beta
AD	=	Alzheimer's Disease
AMP/ADP/ATP	=	Adenosine Mono/Di/Tri-Phoshpate
APOE	=	Apolipoprotein E
APP	=	Amyloid Precursor Protein
BC	=	Betweenness Centrality
BH	=	Benjamini-Hochberg
CC	=	Closeness Centrality
cDNA	=	complementary-Deoxyribonucleic Acid
C,S,D	=	Conserved, Specific, Differentiated
CSF	=	Cerebrospinal Fluid
CNS	=	Central Nervous System
DEA	=	Differential Expression Analysis
DEGs	=	Differentially Expressed Genes
DCGs	=	Differentially Co-expressed Genes
GO	=	Gene Ontology
FAD	=	Familial Alzheimer's Disease
FC	=	Fold Change
FDR	=	False Discovery Rate
FE	=	Fold Enrichment
H	=	Node Homogeneity
IRS	=	Insulin Receptor Substrate
k	=	Node degree
KEGG	=	Kyoto Encyclopedia of Genes and Genomes
LPA	=	Lysophosphatidic Acid
mRNA	=	Messenger-Ribonucleic Acid
NAD	=	Nicotinamide Adenine Dinucleotide
NFTs	=	Neurofibrillary Tangles
N, M	=	Number of nodes, links
NGS	=	Next-generation Sequencing
NSF	=	N-ethylmaleimide-Sensitive Factor
PA	=	Preferential Attachment
PET	=	Positron Emission Tomography
PG	=	Phosphatidylglycerol
PPIs	=	Protein-Protein Interactions
Q	=	Global modularity score
RBP	=	RNA-Binding Protein
SNARE	=	Soluble NSF Attachment Protein Receptor
TF	=	Transcription Factor

Chapter 1

Introduction

"We will never understand complex systems unless we develop a deep understanding of the networks behind them."

- Albert-László Barabási

Complex systems are everywhere around us, even if we realize it or not. Your ability to comprehend what you are reading right now comes from the system of billions of neurons communicating in the brain. It is difficult to derive the total behavior of complex systems solely from knowledge of the individual components. This has led to a paradigm shift in biology from the traditional reductionism to holism in the last century [6]. "The whole is greater than the sum of its parts"[1] is the general idea behind the relatively new field called *Systems biology*. It is about studying the *emergent properties* of biological systems, those that arise from the interactions of the components of a system [6].

Network science has emerged in the 21st century as a response to the challenge of describing complex systems. It is an enabling platform with a wide range of applications in all fields of science; it can be used to study everything from information systems like the World Wide Web, to transportation networks, to social networks and biological networks [7]. There are virtually endless possibilities for what a network can represent. What is especially interesting (and surprising) is that all of these seemingly different systems have fundamental principles in common and can therefore be analyzed with a common set of network parameters. The universality of networks allows us to represent systems of any size, directly visible like social relationships or microscopic like molecular interactions.

Modeling biological systems as networks lets us study them as a whole, focusing on the

[1] supposedly stated by Aristotle in Ancient Greece

emergent properties that would not be observable otherwise. Not only can it provide information about the structure - topology - of the network, but it can elucidate underlying principles of biological processes. It is a simple representation, but has proven effective to capture the properties and behavior of many complex biological systems [7, 8]. However, we are bound to lose some information when we make abstractions of natural systems into simplified models [9], and this is important to consider.

Life itself is dependent on the coherent interactions between thousands of genes and metabolites within our cells [7]. The cells in our body communicate and collaborate in order to adapt to continually changing environments. Each cell plays a role in an extensive network of cells, forming specific tissues, organs, and organ systems. Almost all cells in an organism have the same set of genes, but which ones are active (expressed) determine their particular function. Also, there are many ways to regulate the steps from active genes to translated function, and the complexity generally increases downstream from gene to product. The central dogma of molecular biology is that genetic information flows from DNA to mRNA (transcription) and from mRNA to polypeptides (translation) [10]. Simply put; DNA stores all genetic information, RNA carries and translates the information to make proteins, which then perform a wide range of different functions in the body. This process of *gene expression* is in reality much more complex, but it can be useful to generalize in order to obtain a more systems-level understanding.

Gene expression can be examined on different levels, targeting individual genes, or screening thousands of genes simultaneously. The latter, most relevant for this thesis, is called *gene expression profiling*, or *transcriptomics* (when measuring the whole transcriptome[2]). By studying the activity of thousands of genes at once, one can obtain a global picture of the state of the cell or tissue. It allows for detecting system-level trends that would not be discerned by targeting individual genes. The goal of gene expression analysis is usually to find out more about the function of genes and the regulation of their expression in a context-specific manner. It is essential for understanding normal cellular function, but also what goes wrong at the molecular level in disease development. A number of diseases, such as several cancer types and neurodegenerative diseases, have shown to have defects in the cellular machinery that regulates gene expression [11, 12].

Alzheimer's disease (AD) is the most common form of dementia and affects about 30 million people worldwide [13]. AD is a brain disease with devastating consequences, including neurological function deficits like memory loss and incapacity to complete simple daily tasks. The prevalence in aging populations is expected to increase as people are living longer, yet we have no cure or effective treatment. Even though we know some of the main characteristics of AD progression, much remains unclear. This is partly due to the immense complexity of the human brain, of which we lack a detailed map of nerve cell connections [7]. Further research to gain biological insights is therefore highly needed. For that reason, the World Health Organization (WHO) has promoted it as a public health priority by initiating the Global Action Plan on Dementia

[2] total amount of all RNA transcripts in a cell or tissue in a given moment

2017-2025 [1]. One of the plan's objectives is to increase the amount of global dementia research, and this thesis will contribute in that regard.

The genetic complexity of AD requires it to be studied on many fronts in order to ultimately find a cure or better treatments than those already available. Research in fields such as neuroscience, network medicine, and genetics will be necessary for increased knowledge into the underlying pathogenesis, which offers the hope of developing treatments with therapeutic success [14]. Advances in the global monitoring of gene expression have enabled a system-level study of gene correlations. It is essential to understand how genes and proteins interact with each other and the environment to fulfill their roles and functions [2]. High-throughput transcriptomics data combined with mathematical models to build gene co-expression networks can give vital new insights into complex diseases, such as AD [15]. The CSD method is a newly developed framework at the Department of Biotechnology and Food Science at the Norwegian University of Science and Technology (NTNU) [3]. Recent studies have successfully predicted patterns involved in disease transformation utilizing this method [16, 17]. It improves already existing methods for differential co-expression analysis by including three different types of co-expression: conserved (C), specific (S), and differentiated (D). Therefore, this method was chosen as a basis in this thesis for gaining insights into the development of AD.

The main aim of this thesis was to identify genes and biological processes that have potential roles in the pathogenesis of AD through *system*-level network analysis. Publicly available gene expression data from postmortem brain tissue was used to a perform pairwise comparison of gene correlation patterns in healthy versus disease states. The CSD framework [3] was used for the generation of a differential co-expression network from the transcriptomic data. Analysis of this network was complemented by differential expression analysis (DEA) and protein-protein interactions (PPIs) to gain a better understanding of the molecular interactions underlying complex disease mechanisms. The goal of this integrative analysis was to extract new information not captured by the CSD method alone. Overall, data mining was used as an underlying approach for this thesis. As opposed to testing a specific hypothesis, expression profiling can help identify candidate hypotheses for future experiments.

Chapter 2

Theoretical background

This chapter will give an overview of important theory and concepts for understanding the methodology and analysis performed in this thesis. First, a description of the characteristics and prevalence of Alzheimer's disease will be provided. Then, an introduction to network theory will be given, with focus on the concepts specifically relevant for this thesis. For more detailed information, the interested reader is encouraged to read Barabási's book of Network Science [7]. Further, gene expression profiling and analysis through network construction will be explored. An extensive literature search was done to provide a foundation of the research already performed and the future work needed. The methodology used in this thesis is based on the CSD framework of Voigt et. al. [3], which will be described in section 2.4.1. Finally, a brief introduction to PPIs is made.

2.1 Alzheimer's disease

Alzheimer's disease (AD) is a progressive neurodegenerative brain disease[1] that results in the loss of cognitive functions [18]. Common symptoms include short-term memory loss, confusion in familiar places, problems with finding words and behavioural changes, leading to a reduced ability to perform everyday tasks. AD is the most common cause of dementia, and 29.8 million people worldwide (2015) are estimated to have the disease, mostly people older than 65 [13]. Age is the biggest risk factor, and prevalence is expected to increase rapidly as the world population's life expectancy rises [19]. Although AD mainly affects older people, it is not an assured consequence of aging. Further, around 5 % of the cases are early-onset AD, starting in people younger than 65 years [19]. The disease progresses through gradually worsening symptoms, often resulting in a total dependence on others for personal care and the inability to recognize

[1] Neurodegenerative diseases is a group of diseases which show loss of function and/or death of nerve cells in the central- or peripheral nervous system (CNS or PNS).

family and friends. It has a poor prognosis, with a life expectancy after diagnosis of only three to seven years [20].

AD is a complex brain disease, and its course of action is associated with several biological mechanisms. The initial cause of AD is poorly understood, and it has no known specific trigger. Still, there are many pathological features associated with the disease that have been well studied. The two major hallmarks of AD are related to abnormal protein aggregation; i) amyloid β (Aβ)-containing senile plaques and ii) hyperphosphorylated tau-containing neurofibrillary tangles (NFTs) [18]. AD is therefore classified as a proteopathy, a disease associated with aggregation of misfolded proteins. Aβ is a 40-42 amino acid peptide generated by proteolytic cleavage of Amyloid Precursor Protein (APP) by γ- and β-secretases [21]. The extracellular build up of amyloid oligomers (2-12 peptides) and plaques (larger aggregates) has a toxic effect by blocking cell-to-cell signaling at synapses [14]. It also seems to trigger immune reactions that cause the destruction of disabled nerve cells by programmed cell death [14]. NFTs form *within* nerve cells by abnormal aggregation of the microtubule[2]-associated protein *tau* [14]. This protein normally stabilizes the microtubules, but in AD it becomes hyperphosphorylated and aggregates into insoluble threads (tangles). This leads to microtubule disassembly, which obstructs nutrients from reaching the cells - eventually resulting in cell death [14]. Other features associated with AD include oxidative stress, mitochondrial dysfunction, inflammatory responses, aberrant signaling and lipid metabolism, and DNA damage [18]. These can both precede or be a consequence of protein aggregation, but the underlying mechanisms remain elusive [12, 5].

The symptoms of AD is ultimately a result of losing nerve cells or some of their synapse connections to other cells [22]. This leads to brain tissue damage and the shrinking of the brain at the macroscopic level [14, 19]. Synaptic loss and neuronal death leads to cognitive decline specific for the brain region affected. Although multiple areas are affected, it is commonly understood that protein aggregation starts in the brain region called hippocampus, responsible for storing memories [23]. Short-term memory loss is therefore one of the earliest symptoms. Proteins then progressively invade other parts of the brain. Braak staging is used to characterize the severity of brain damage associated with NFT evolution [24]. In the six stages (I-IV) the tau aggregation spreads progressively into different parts of the human brain. It is however important to note that the process is a continuum where the stages can overlap [14, 19].

The genetic basis behind AD is heterogeneous - it is likely that the interplay between several genetic changes plays a role in disease development. In addition to dividing AD into early- and late-onset, the disease can be categorized by heredity. Most AD cases are sporadic, meaning that they occur in people with no history of AD in their family [14]. Rare cases are familial (FAD), where the inheritance appears to be autosomal dominant [14]. This means that each child of an individual with the disease has a 50 % chance of inheriting the pathogenic gene variant. This form is predominantly early-onset, and the earlier the onset of AD, the more likely there is a genetic cause [14]. Four genes have to date been linked to cause FAD. Mutations in *APP, Presenilin 1 (PSEN1),* or *Presenilin*

[2]Microtubules are long, tubular structures part of the cytoskeleton, important for transport of nutrients and other molecules.

2 (PSEN2), are each causative of early-onset FAD, while the ϵ4 allele of *Apolipoprotein E (APOE)* is mainly a risk factor of late-onset FAD [25]. On the other hand, the cause of the most common type of AD (sporadic) is unknown, but genetic and environmental interactions are likely to play essential roles. It is a complex disorder involving multiple susceptibility genes [14]. Many genes have already been associated with increased risk of AD, and there are lots of ongoing research on this. A list of 499 genes associated with AD and their relevance scores from MalaCards database [25] can be accessed with this doi: 10.6084/m9.figshare [26].

There are several risk factors other than genetic factors that can influence disease progression, both non-modifiable like age and modifiable such as lifestyle (sleep, diet, exercise) [27]. Studies have shown that the risk of dementia can be reduced by exercising regularly, eating healthy, not smoking, reducing the consumption of alcohol, and maintaining a balanced blood pressure, cholesterol and blood glucose levels [28, 29]. It has also been demonstrated that higher intelligence and educational levels is associated with a reduced risk of developing AD [30, 31]. This is due to the higher cognitive reserve - greater resilience against brain damage [31]. By being engaged in mentally stimulating activities, the abundance and redundancy of synapses (neural connections) is increased. This is possible because of neural plasticity, the brain's ability to change synapses based on experience [32].

The diagnosis of AD is complicated and usually requires a comprehensive assessment. There is no single test for determining if someone has AD, but various approaches and tools have been developed [33]. Diagnosis usually relies on the doctor spending time with the patient, checking for signs and symptoms and taking their medical history. Testing the mental status by neurological examinations is crucial in diagnosing AD. Other assessments may include blood tests or brain scans - the latter is usually done to rule out other conditions that produce similar symptoms, such as tumor or stroke [33]. Overall, it can be difficult to diagnose AD, partly because the disease can be considered a continuum. From initial neuronal damage to clinical symptoms are detectable can take many years [27]. Identifying the disease in the preclinical stage (before symptoms occur) is now a major research focus [27]. Advanced techniques[3] are available that can detect Aβ and tau biomarkers in the brain at a preclinical stage, but these are invasive and expensive [27]. It is proposed that a detection of both pathological hallmarks can be used to define AD, even in the absence of cognitive symptoms [27]. Currently, however, a diagnosis is only definitely confirmed by brain autopsy after death [33].

At the moment there is no treatment available that can cure or alter the course of AD. However, many researchers and drug companies are working on the development of drugs targeting the disease. The main ongoing therapeutic approaches are targeting the protein aggregation process, either by preventing the formation or misfolding/aggregation of the disease-causing proteins, or by promoting their removal [34]. A more detailed understanding of how protein aggregation connects to tissue degeneration is needed to develop successful therapies, and this will likely involve systems biology and network medicine.

[3] e.g. positron emission tomography (PET)-imaging and cerebrospinal fluid (CSF) sampling

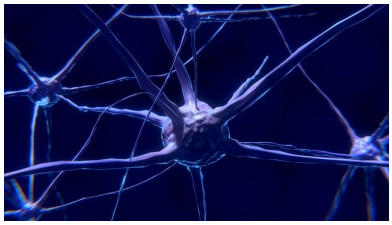
2.2 Network Theory

In this section some of the basic properties of a network will be mentioned, using biological networks to demonstrate. The way concepts are explained is largely based upon the "Network Science"-textbook [7].

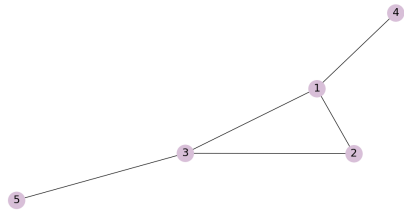
A network is a representation of a (real) system of components - called *nodes*, and how they connect with each other by *links* [7]. When explaining mathematical properties of a system, which is not necessarily modeling specific real relationships, it is conventionally called a *graph*, with the objects being called *vertices* and the interactions called *edges*. Distinctions are made whenever it is appropriate, but in most cases these terms are used interchangeably. The most basic characteristics of a network is the number of nodes (network size), N, and the number of links, M [7]. The links can be either undirected (straight lines) or directed (arrows). An essential aspect of network theory is connectedness, describing how well the nodes in the network connect with each other overall. A network can consist of one or multiple *connected components*, which are subsets of nodes for which every pair of nodes is connected by at least one path. The largest connected component, given that its size is substantially larger than other components of the network, is often referred to as a *giant component*. An edge whose deletion increases the number of connected components may be called a *bridge* [35].

Biological networks can be defined at different levels, with the system being for example complete cells or a set of interacting biomolecules. The neural network is a representation of connections between billions of nerve cells in the brain [7]. It can give us information of how the brain works in order to maintain cognitive functions and how it is affected by disease. Figure 2.1 shows a small example of a neural network, where the nerve cells are represented by nodes and their synaptic connections by links. For simplicity only 5 nodes are included, representing only an infinitely small fraction of all the nerve cells in the human brain (system). Disease progress could be modelled by the changes in size (N) and connectivity of the network. By using the network as a model of disease progression, one could study neurodegeneration as node removal and neuronal plasticity as addition/removal of links. Node removal could represent the death of a neuron, while link removal could represent the case of synaptic loss. In this thesis, the reader can imagine that we move further into the nerve cells and take a look at the associations among genes (section 2.4) and potential interactions between the proteins they encode (section 2.5). These interactions are in essence responsible for the form and function of the nerve cells, and studying networks of such interactions can give valuable new information.

There are many different ways to visualize the same network, so in order to obtain a precise and unique description we need to use the language of mathematics. In addition, many networks of interest are far too large and complex to visualize, and it is essential with mathematical modeling and computational power to extract meaningful information from them. In the following subsections, we will provide a brief overview of some common properties, both local and global, used to analyze complex networks.



(a) Brain neurons [CC0 1.0]



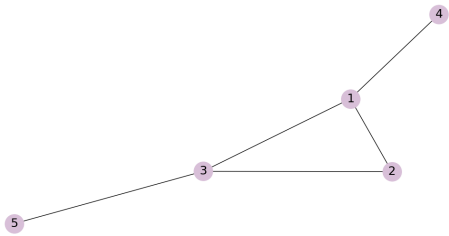
(b) Network representation

Figure 2.1: Example of a biological system and network representation. **a)** Brain neurons connected by synapses. Image from [36], distributed under the Create Commons licence CC0 1.0. **b)** Network representation of neurons, created in Python using NetworkX and Matplotlib. N (nodes), M (links) = 5.

2.2.1 Adjacency matrix and node degree

In addition to representing networks graphically, one can represent them mathematically through *matrices*, which are more useful for analysis purposes [37]. For *simple graphs* (no self-loops or multiple edges), the adjacency matrix A is a square nxn-matrix, where n is the number of nodes and each element a_{ij} quantifies the connection between nodes i and j [37]. The adjacency matrices of *unweighted* networks have binary values; 1 if the nodes are connected by an edge, and 0 if a lack thereof (Fig. 2.2). *Weighted* networks, on the other hand, have adjacency matrices where the elements take on a continuous range of numbers representing the weights of the edges. For *undirected* networks, the adjacency matrix is symmetrical, meaning that $a_{ij} = a_{ji}$ (Fig. 2.2). A correlation network is an example of this, since if one node correlates with another node it necessarily implies that the opposite is true. This is not the case for *directed* networks, such as regulatory networks, where the connection between two nodes mean that one is *regulating* and the other is *being regulated*. Then the row and column of the adjacency matrix would each represent one direction of interaction. Overall, the adjacency matrix is a simple illustration of network topology.

The matrix representation is a compact way to store information which permits us to calculate common network properties using basic concepts from linear algebra [38]. One of the most fundamental properties of a node i is its degree k_i, which is the number of edges adjacent to the node [8]. This can be calculated, for undirected networks, by summing over the elements in its respective row or column in the adjacency matrix[37]. For example, the degree of node *one* in Fig 2.2 is $k_1 = 3$, which is the sum of either the first row or first column. The degree is thus equivalent to the number of neighbors the node has, assuming that the network is without self-loops and multiple edges [8]. The word *degree* should not be mistaken with the word *connectivity*, which is related to the number of nodes or edges whose removal is necessary to disconnect a graph. For directed networks, one distinguishes between in-degree ki_{in} - the number of arrows (edges) pointing towards i - and out-degree ki_{out} - the number of arrows pointing away from node i [7]. From now on, weighted and/or directed networks will be omitted from discussion unless explicitly stated otherwise, as they are outside the scope of this thesis.



$$A_{ij} = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Figure 2.2: The same network as in Fig 2.1 with its corresponding adjacency matrix A_{ij}. The network is undirected and unweighted, seen in the matrix by symmetry ($a_{ij} = a_{ji}$) and binary values, respectively.

2.2.2 Degree Distribution and Scale-Free Networks

By considering all node degrees in a network, a global parameter called the degree distribution $P(k)$ can be obtained. This is a probability distribution, which gives the fraction of nodes with degree k in the network [7]. The nature of this distribution has an impact on the general structure of the network. For random networks (such as Erdos–Renyi), most nodes have an average number of neighbors, which leads to a bell-shaped degree distribution [7]. The average degree in a network of N nodes and M edges is $< k >= 2M/N$ [7]. However, for real networks, including most biological networks, the average degree is not necessarily a good representation of the overall network structure, as there is no "typical" degree. This lack of a general scale is why these types of networks are called *scale-free*. This network class has degree distributions that resembles a power law,

$$P(k) \sim k^{-\gamma} \tag{2.1}$$

where the *degree exponent* is typically observed in the range $2 < \gamma < 3$ [39]. The exponent is the slope of the line that fits the data on log-log scale. In scale-free networks, most nodes thus have a low degree, while only a few nodes are highly connected to many other nodes. The highest-degree nodes are called *hubs*, and the existence of hubs is a characteristic of such networks [7]. Another characteristic of scale-free networks is their robustness to random failures, as removing random nodes or links are not likely to affect the whole structure substantially. Conversely, they are very vulnerable to targeted attacks: removing the hubs can rapidly disconnect and break down the whole system. For this reason, defective hubs might contribute to more of the dysregulated processes associated with disease progression than low-degree nodes. Therefore, the study of hubs in a network is generally an essential step in network analysis.

Interestingly, the scale-free property suggests that biological networks are not organized randomly. There is a general belief that scale-free networks grow because of preferential attachment (PA). This growth mechanism is based on the preferential addition of new nodes to already highly connected nodes [39]. Even though this growth mechanism leads to a scale-free degree distribution, the converse is not necessarily true -

not all power law distributions are generated from PA. It is important to notice that this scale-free degree distribution is more a behaviour than a consistent definition, and that it is not always present in the whole range of degrees. Even though it has been reported to appear frequently in many different types of real-world networks, some have recently argued that scale-free networks are actually rare [40]. This is controversial and further work is needed to assess the universality of this property and potentially discover novel more realistic degree structures in networks. Going beyond the degree distribution, we find a somewhat related network measure called degree correlations that reflects the way nodes connect to each other, which is not captured by $P(k)$ alone.

2.2.3 Degree correlations

Mixing patterns in a network can influence the overall behaviour of the system. *Assortative mixing* is an important network property, which describes node correlations, most often in terms of degree. Degree correlations capture how nodes with particular degrees interact with each other. In *assortative* networks, nodes with high degree have a tendency to connect to other nodes with high degree (hubs associate with hubs), while networks in which high-degree nodes tend to pair up with low-degree nodes are called *disassortative* [41]. In the latter, nodes with similar degree seem to repel each other. If no correlation is found between the degree of adjacent nodes, the network can be called *neutral*.

One common way to (qualitatively) determine node degree correlations is through the *neighborhood connectivity distribution* $k_{nn}(k)$, which is the average degree of the nearest neighbors of a node with degree k as a function of the degree itself [7]. The neighborhood connectivity, along with its approximation is shown in equation (2.2):

$$k_{nn}(k) = \frac{1}{k_i} \sum_{j=1}^{N} A_{ij} k_j \approx a k^{\mu} \tag{2.2}$$

where *nn* stands for "nearest neighbor", a is a constant and μ is the degree correlation exponent [7]. The sign of the correlation exponent (slope) reveals if the network is assortative ($\mu > 0$), neutral ($\mu = 0$) or dissassortative ($\mu < 0$).

Most real networks display some form of degree correlations. Social networks are known to have an assortative nature, where highly connected people tend to know each other. The fact that celebrities often date other celebrities, is not random. On the other hand, most technological (World Wide Web, Internet) and biological (metabolic, protein interaction) networks are disassortative [42].

Degree correlations are important for many reasons, not only for academic purposes. They influence other network properties such as clustering, shortest paths, diameter, and its robustness to perturbations [7]. More information on these network parameters can be found in the "Network science"-textbook [7].

2.2.4 Clustering

Clustering is an important topological concept in network theory, which provides information on neighborhood relationships not captured by the degree itself. The *clustering coefficient* is a node parameter measuring how connected the neighbors of a node are to each other. It is defined as the ratio of existing links connecting a node's neighbors to each other relative to the maximum possible number that could exist between them [7]. For undirected networks it can be calculated by:

$$C_i = \frac{2E_i}{k_i(k_i-1)}, \tag{2.3}$$

where E_i represent the number of links between the k_i neighbors of node i [7]. The clustering coefficient has a value between 0 and 1, where none or all of the neighbors of the node link to each other, respectively. In the example network in Fig. 2.2 node *one* has a clustering coefficient of $C_1 = \frac{2\cdot1}{3\cdot2} = 0.33$. The extent of clustering - triangle formations - in the entire network can be captured by averaging the clustering coefficients for all nodes.

2.2.5 Centrality measures

There are other ways than degree centrality to describe how important a node is in a network. Closeness- and betweenness centrality are two related centrality measures that are both based on distance, more specifically the shortest paths. The *shortest path* (*d*) is defined as the path with the minimal number of edges between two nodes [7].

Closeness centrality (CC) is a measure of how close a node is to all other nodes in the network. It is defined as the inverse of the sum of all the shortest paths from i to all other nodes in a connected component (Eq. 2.4) [7]. A node with high CC is in a central position where it can rapidly spread information to all other nodes. In the earlier network example (Fig 2.2) node *one* and *three* are equally close to the other nodes and both have $CC = 0.8$.

$$CC_i = \frac{1}{\sum_{j\neq i} d_g(i,j)}, \tag{2.4}$$

Betweenness centrality (BC) is a measure of how often a node is a bridge between other nodes. It is defined as the fraction of shortest paths that go through node *i*:

$$BC_i = \sum_{j,k=1;j\neq k\neq i)}^{N} \frac{\sigma_i(j,k)}{\sigma(j,k)}, \tag{2.5}$$

where $\sigma_i(j,k)$ is the number of shortest paths between nodes j and k that pass through node i, and $\sigma(j,k)$ is the total number of shortest paths between nodes j and k [7]. A node with high BC has an important role of information transfer between different parts of the network. Node *one* and *three* in Fig 2.2 function as "bridge" nodes in the small

network with $BC_{1,3} = 0.5$. All paths from the other nodes (with BC = 0) must go through these central nodes.

2.2.6 Communities and modularity

The identification of network communities (often called modules in relation to gene co-expression networks) is a common approach in network analysis that can provide insight into functional properties of locally dense neighborhoods. The general assumption is that nodes forming part of the same topological module have closely related functions. For example, disease genes have been shown to have a tendency to interact and form disease modules, which can aid in the prediction of disease pathways and other disease genes [43].

In network science, a *community* is a group of nodes with a higher likelihood of connecting to each other than to other nodes of the network [7]. The central idea is that nodes are assigned to subgraphs based on the link structure of the network. There are several ways to define communities, but once clearly defined, we could identify them by assessing all possible partitions until we have found the one that best fits the definition. Yet, inspecting all partitions is computationally infeasible as the number of partitions grows faster than exponentially with the size of the network [7]. Due to this we need heuristic algorithms, where the common goal is to optimize a score called *modularity*. The global modularity score, Q, measures the quality of partitioning a network into n_c communities [7]. It can be calculated by Eq. (2.6),

$$Q = \sum_{c=1}^{n_c} \left[\frac{M_c}{M} - \left(\frac{k_c}{2M} \right)^2 \right] \tag{2.6}$$

where M_c is the total number of links within a community and k_c is the total degree of the nodes in the community [7]. The value is positive when there are more links within groups than expected by chance. The higher Q the better the community structure of a partition, up to a maximum of $Q = 1$. [7].

Many different algorithms exist for partitioning a network into smaller modules. Overall, they can be divided into agglomerative (bottom-up) and divisive (top-down) methods. Either may assign a unique group or multiple groups to each node. In this thesis, the Louvain community detection algorithm was chosen because it is a widespread and fast algorithm that can easily be implemented in Python with the NetworkX package [44]. The Louvain-method is an agglomerative algorithm that maximizes modularity in a two-step iterative process [45]. Initially, all nodes are assigned to their own unique community. The first step is a local modularity optimization phase: each node is moved to the community of a neighbor that leads to the largest positive change in the modularity. This is repeated for all nodes until no further improvements in modularity is achieved. In the second phase, nodes of the same community are joined to build a new network whose nodes are the communities. The steps are repeated iteratively until maximum modularity is reached. For more details, see the original article [45].

2.3 Gene expression profiling

The general theory on gene expression in this section is primarily based on [46, 10]. *Gene expression* is the process of going from genetic information to a functional product. The product is usually a protein, but it can also be ribonucleic acid (RNA). Cells regulate their gene expression levels as a response to different environmental signals. All (somatic) cells in our body have the same set of genes, but they still turn into entirely different cell types with specific morphologies and functions. This is a result of the process of *differentiation*, where cells become specialized. Specific genes are expressed based on what is most appropriate in a particular time and condition. For instance, skin cells and nerve cells "turn on" different genes, which is the main reason why they look and behave differently. Also, different cells of the same type may have different expression patterns depending on their external and internal state. Molecular pathways convert environmental signals - generally through a cascade of events - into a change in gene expression. Regulating which genes are active and at what level is a complex process that can happen at all the different levels of expression and involve several different molecules along the way. In general, the amount of protein (or other functional products) at a given moment is the difference between what is synthesized and what is being degraded (not considering cell export). Maintaining a balance between these two processes is important for cell efficiency. In order to limit energy waste on synthesizing proteins that are not needed in a particular moment, gene expression is most often regulated at the level of transcription. Some proteins, called transcription factors (TF), are able to regulate whether other genes are activated or repressed. Together with other TFs, they provide a combinatorial effect that contributes to determining the number of RNA transcripts made from a gene. Transcriptomics techniques can measure the whole *transcriptome*, meaning all RNA transcripts available in a specific context - both mRNA encoding proteins, and other types of non-coding RNA (ncRNA) that might have regulatory roles as RNA directly.

Several techniques exist for the global monitoring of gene expression, with DNA microarray and RNA sequencing (RNA-seq) being the two most widely used [47]. Both measure large-scale RNA expression, but which technique to use depends on several factors. It can for example depend on what genome information is available for the species of interest, which data analyses to use and often most importantly, it is a balance between cost and performance. DNA microarray, often referred to as just microarray, was developed first and is well established in research. RNA-seq, on the other hand, is a more newly developed technology, based on next-generation sequencing (NGS) [48]. Microarray has generally been able to generate high-throughput data at a lower cost than RNA-seq, but it is dependent on prior sequence knowledge. Unlike microarrays, RNA-seq does not rely on probes, and can therefore detect structural variations such as alternatively spliced transcripts, and even previously unknown genes [48]. RNA-seq is therefore increasingly a preferred platform to use, but there are still challenges in storing the large amount of data, and standard protocols for downstream analysis are yet to be established [47, 48]. The CSD framework used in this work (explained in 2.4.1) is suitable for both data types, and so the availability of high-quality gene expression data was the deciding factor. The transcriptomic data used in this thesis was taken from a

microarray experiment, and this technique will therefore be explained in more detail. The following information was largely accessed from two books [2, 10].

2.3.1 DNA microarray

DNA microarray is a high-throughput assay for measuring the relative amount of RNA in a sample, usually at the level of the whole transcriptome [10]. It gives an indication of the relative activity of previously identified genes in the particular cell and condition. The microarray technique is based on DNA-DNA hybridization, the binding of complementary sequences. A microarray chip is a solid support (glass slide or nylon membrane) with microscopic arrays containing different DNA segments of known sequence, called probes, which are complementary to all possible mRNA that a cell might express. These are used to recognize and bind complementary sequences in the experimental sample [10].

The total RNA from the experimental sample is first isolated, and then reverse transcribed to cDNA (with either primers to amplify only mRNA or random primers to amplify all RNA). The cDNA is further labeled and hybridized to the arrays according to the experimental strategy. There are many varieties in the experimental design (platform) such as probe type, labelling and detection method. The probes can be either complementary DNA (cDNA) or oligonucleotides (short nuclotide sequences). Probe synthesis can be done prior to deposition or *in situ*, and the attachment to the spots on a chip by robot spotting or photolitography, respectively. The target sample is labelled either with a radioactive isotope or more commonly with a fluorescent dye. Detection can happen for each experimental sample separately, or by mixing the two samples (usually case and control) with distinguishable labels. The former is called single-color or 1-channel detection, in which a single RNA sample is labeled and hybridized alone to the chip. In the latter approach, dual-color or 2-channel, the two differently labelled samples are hybridized together on a single microarray. Both are commonly used, and the overall performance of each is found to be similar [49]. More detailed information about the different types of microarrays can be found in [2].

After washing off nonspecific binding, the hybridized chips are scanned with a laser and the signal extracted from the digital images is analysed and quantified by data software. The observed amount of hybridization detected for a specific probe is proportional to the number of corresponding RNA transcripts present (at the location). Overall, the measured intensities indicate the relative level of gene expression, so the chip provides a snapshot of which genes were actively transcribed at the time and condition when the sample was taken [10].

Finally, in order to compare experimental samples, the resulting data must be normalized and corrected for background noise. Normalization of the measured intensities is important to adjust the differences in starting amount of RNA, and to reduce the bias from systematic variation in the microarray experiment. A variety of nonbiological sources, such as pipetting errors and label-detection efficiencies, can affect the measured expression levels. These need to be eliminated to enhance the reliability of the data in further downstream analyses. There are many different normalization algo-

rithms available, but these will not be described here. Ultimately, and most importantly, data mining can be used to extract biologically relevant information about the system being studied from the large amount of data that the arrays generate. There are many methods available, but they all depend on the integration of biological knowledge with statistics and computer science [2].

2.3.2 Differential gene expression analysis

A common data mining approach in transcriptome profiling is the identification of genes that are differentially expressed between sample groups in the data. Differential expression analysis (DEA) determines the quantitative changes in mean levels of gene expression across conditions [50]. When conditions compared are disease vs control, the change in gene expression can provide clues about the mechanisms involved in the molecular pathogenesis of the specific disease [18]. Being representative of the relative amount of gene transcripts, this change in expression level indicates a transcriptional regulation as a response to the altered cell condition. The increase (upregulation) or decrease (downregulation) of a specific mRNA suggests a changed need for the protein encoded by that mRNA. The change in the abundance of that protein can directly or indirectly affect the rate of a biological pathway, potentially involved in the pathological condition. Misregulation of certain genes can therefore increase the risk of disease or accelerate the progression of disease [51].

Differentially expressed genes (DEGs) between two sample sets are usually found by calculating the *fold change* (FC) and testing for statistical significance [52]. FC is a measure of the ratio between two quantities, the change of one with respect to the other. Even though this is an intuitive measure, it treats increases and decreases in expression levels differently. A gene that is 2-times (doubled) up-regulated has a ratio of 2, whereas a 2-times (halved) down-regulated gene has a ratio of 0.5. Logarithmic ratios are commonly used as they make the ratios symmetrically distributed, which makes it easier to compare up- and down-regulated genes in a similar fashion [50]. The most widely used transformation is logarithm base 2, because it handles numbers and their reciprocals symmetrically [50]. So for the example above $\log_2(2) = 1$ and $\log_2(\frac{1}{2}) = -1$, up- and downregulation by the same factor (2) is the same value with opposite sign. A gene expressed at a constant level (FC = 1), hence not differentially expressed, will have $\log_2(\text{FC}) = 0$. It is also important that the expression values are normalized to inhibit bias [50].

An arbitrary cutoff value of FC (generally twofold) has traditionally been used as a fixed threshold for DEG classification in microarray experiments [50]. However, the fold change is in itself not a statistical test, because it does not provide a confidence level when designating a gene as differentially expressed or not [2]. Therefore, it good practice to use a statistical method such as a two-sample t-test. The *Student's t-test* assesses whether the means of two groups are statistically different from each other, by taking the Standard Error (SE) into account [53]. A t-value for a given gene is calculated by Eq. (2.7), where M is the mean expression value, S is the standard deviation and subscripts represent the two groups to be tested [53].

$$t = \frac{M_1 - M_2}{\sqrt{S_1^2 + S_2^2}} \tag{2.7}$$

It estimates the signal-to-noise ratio, where signal is the observed difference between sample means (numerator) and the noise is the standard error of the difference between the means (denominator). After the test statistic is computed, it is converted to a *p-value*, which represents the probability that the observed difference could have occurred by chance [52]. It is common to use a nominal level of 0.05 as a significance value, below which genes are regarded as significant [52]. Due to the large number of genes (>20,000 in the human genome) to be tested in a transcriptomic study, one would expect a substantial amount of false positives if only individual p-values were considered [2]. For example, if 20 000 genes are tested with 0.05 as significance threshold, then $20000 \times 0.05 = 1000$ genes are expected to be found differentially expressed by chance. It is therefore important to correct for multiple testing.

One common approach to solve for the multiple comparison problem is called the *Benjamini-Hochberg* (BH) method [54]. It considers the false-discovery rate (FDR) - the expected number of false positives among all genes initially recognized as differentially expressed [52]. The FDR-value is limited to a chosen level α, commonly 0.05. This means that 5 % of the "significant" results will be accepted as false positives. The BH method adjusts (enlarges) the original p-values based on their rank, i and the total number of tests, m [54]. First, the p-values are ordered from smallest to largest and assigned ranks. Then the FDR-adjusted p-values, also called *q-values*, are defined recursively beginning with the largest. The largest q-value and the largest p-value are the same. The rest of the q-values are calculated by

$$q_i = p_i * \frac{m}{i} \tag{2.8}$$

and compared to the previous q-value. The smaller value is kept as the adjusted p-value. Finally, all tests with q-values less than or equal to the chosen α (FDR) are considered significant [54].

2.3.3 Gene Co-expression analysis

While analyzing the differential expression of individual genes can predict their biological function, it does not tell us about how genes may interact among each other. In fact, biological molecules rarely act alone, and this limitation can be addressed by co-expression analysis [47]. Two genes are said to be *co-expressed* if their gene expression levels have a similar pattern across samples, due to the amounts of RNA transcripts rising and falling in a concordant fashion [55]. It has been demonstrated that co-expressed genes have a tendency to be functionally related or have underlying regulatory relationships [56, 57]. They might encode proteins that are part of the same pathway or protein complex, or that are regulated by the same transcriptional program [58]. Studying co-expression patterns can therefore provide insight into the underlying biological processes [56]. It is an essential tool for the functional annotation of unknown genes based

on the "guilt by association"-principle [55]. Based on the assumption that co-expressed genes are functionally related, it can be used to predict the function of genes within the same co-expression module [56], and to identify disease gene candidates neighboring genes already associated with a certain disease [59, 60].

Correlation

In order to study the coexpression of genes it is important to define a measure that quantifies the similarity between expression profiles. *Correlation* is a similarity measure that is commonly used to determine whether two genes have similar expression patterns [61]. Generally speaking, correlation is a statistical measure of relatedness between two or more variables. It can be used to indicate predictive relationships, but the presence of a correlation does not necessarily imply causation (in either direction). If any, the causes underlying the correlation may be indirect or unknown. It does however provide *possible* causal relationships that can be interesting to investigate further [60].

Several types of correlation coefficients exist, with the Pearson's correlation and Spearman's rank correlation being among the most common [62]. They both measure the strength and direction of the association between two variables. The value varies from -1 to +1, where ±1 indicates a perfect correlation, and values becoming weaker as they approach zero. The sign indicates the direction; a + sign meaning positive relationship and a - sign indicating a negative relationship. The two correlation coefficients differ in the type of relationship they infer. Whereas Pearson's is only sensitive to linear relationships, Spearman's was developed with increased robustness in identifying nonlinear relationships [62]. As gene pair correlations are not necessarily linear, the Spearman's correlation coefficient was used as the similarity measure in this thesis.

As the name implies, the Spearman's rank correlation finds the dependence between the *rankings* of two variables [63]. It is therefore more robust to outliers, since the extreme values of the raw data are not used directly but turned into ordered values. It is a non-parametric test that can be used without the need for any assumptions of data distribution[63]. It does require that the data contains paired samples, which means that the two variables must have the same number of measurements. The Spearman's correlation, *rho*, can be calculated by:

$$\rho = 1 - \frac{6\sum d^2}{n^2(n-1)} \tag{2.9}$$

where n is the number of measurements and d is the difference between the ranks of the corresponding variables [63]. The Spearman's coefficient looks for monotonic relationships between the ranked variables, which can happen in two ways [63]:

1. as the value of one variable increases, the other variable never decreases (monotonically increasing), or,

2. as the value of one variable increases, the other variable never increases (monotonically decreasing).

2.4 Differential Gene Co-expression Networks

After correlation has been computed for each possible pair of genes in the gene expression data, a network can be constructed where nodes represent genes and links represent the pairwise correlation [60]. Genes are connected by an undirected link if there is a significant association between their expression levels across samples. Such co-expression networks enable scientists to illustrate gene correlations in a graphical way at a genome level. It is an holistic approach for analyzing high-throughput transcriptomics data. As an extension of gene co-expression, it is increasingly common to study *differential* gene co-expression, comparing the co-expression patterns between two conditions, such as disease states, tissue types or organisms [60]. As the word "differential" indicates, the goal is to find differences in co-expression patterns in order to discover processes specifically relevant to a certain condition [3]. This can give better conclusions about transcriptional regulation than the differential mean expression alone [9].

(Differential) gene co-expression networks can be analyzed by the many existing network algorithms, some of which were mentioned in section 2.2. However, it is important to note that it is not a trivial task to infer biological relationships from these correlation networks. As the networks are undirected, a link between genes is not evidence of a causal relationship; it merely represents a coinciding expression pattern [3]. Genes that are simultaneously active genes can indicate that they are active in the same biological process, but correlation does not distinguish between regulatory and regulated genes [60]. In this sense it is different from directed networks such as gene regulatory networks. Two co-expressed genes might be correlated due to different relationships: i) direct effects, such as gene A causing gene B or vice versa, or bidirectional causation, ii) indirect effects (transitivity), or iii) confounding effects (common regulator), such as nutrient availability affecting both or a TF regulating both genes [9]. Despite the difficulty in determining the (potential) causal relationship, varying correlation patterns hint at condition-mediated regions of the network where it could be worth conducting further detailed analysis. Differential gene co-expression networks have therefore gained major interest, especially during the last decade. Many methods and tools for studying differential co-expression networks have been reported, and the relevance of each depends on the research question.

2.4.1 The CSD Framework

In this thesis, the CSD framework will form the basis for differential co-expression network construction [3]. It was developed by André Voigt and Eivind Almaas at the Department of Biotechnology and Food Science at NTNU. A systematic comparison between the CSD method and other available methods was presented in their work. They found that the predictive power of the CSD method is higher than that of any of the other nine differential co-expression methods studied for comparison [3].

What is new in this CSD framework is that it distinguishes between three different ways a change in gene pair correlations can happen when comparing two conditions (or tis-

sues or organisms) [3]. The name *CSD* corresponds to the three types of co-expression; *conserved* (C), *specific* (S) and *differentiated* (D). Conserved (C) means that there is a similar co-expression pattern between the gene pair, strong and same-sign correlation, in both conditions. Specific (S) refers to cases where one of the conditions show strong correlation of any sign between the genes, while in the other condition there is a weak or no correlation between the gene pair. Finally, differentiated (D) co-expression represents strong gene pair correlations in both conditions, but with oppositely signed values [3].

The CSD framework includes a set of software programs for network generation, where the nodes are genes and the edges between gene pairs represent one of the three types of co-expression relationships. The method starts with gene expression profiles from two different datasets, obtained for example by microarray or RNA-seq (as described in section 2.3). The first step towards a differential co-expression network is to define a similarity measure, commonly a correlation parameter. In this case, the Spearman's rank correlation coefficient is used as default to calculate pairwise correlations across all genes in each dataset separately. The correlation coefficient $\rho_{ij,k}$ represents the co-expression of gene pair (i,j) in condition k. The values of $\rho_{ij,k}$ range from -1 to 1, where values close to the bounds indicate strong correlations, whereas values close or equal to zero indicate weak/no correlation. Each pairwise correlation gives a similarity score s_{ij}, resulting in a similarity matrix $S = S_{ij}$.

The next step is to turn the co-expression values into *differential* co-expression values by calculating the *change* between conditions, using equations (2.10)-(2.12);

$$C_{ij} = \frac{|\rho_{ij,1} + \rho_{ij,2}|}{\sqrt{\sigma^2_{ij,1} + \sigma^2_{ij,2}}} \tag{2.10}$$

$$S_{ij} = \frac{||\rho_{ij,1}| - |\rho_{ij,2}||}{\sqrt{\sigma^2_{ij,1} + \sigma^2_{ij,2}}} \tag{2.11}$$

$$D_{ij} = \frac{|\rho_{ij,1}| + |\rho_{ij,2}| - |\rho_{ij,1} + \rho_{ij,2}|}{\sqrt{\sigma^2_{ij,1} + \sigma^2_{ij,2}}} \tag{2.12}$$

where C_{ij}, S_{ij} and D_{ij} represents the C, S and D relationship scores between a pair of genes i and j, with the numbers in subscript representing different conditions, tissues or organisms [3]. The numerators are absolute correlations, while the denominators are the root of the summed correlation variance. This variance can be estimated through a sub-sampling algorithm, although as not part of the scope of this thesis, the interested reader is referred to the work of Voigt et al. [3]. The relationship scores lie in the range $[0, \infty]$, and are designed to increase as the given differential co-expression gets stronger. Figure 2.3 illustrates where these three possible differential co-expression relationships lie on a plot where the axes are the gene pair correlations. The scores have large values within their respective colored regions (blue = conserved, green = specific, red = differentiated), and the white area represents combinations of correlations that are too weak to be included in the final network.

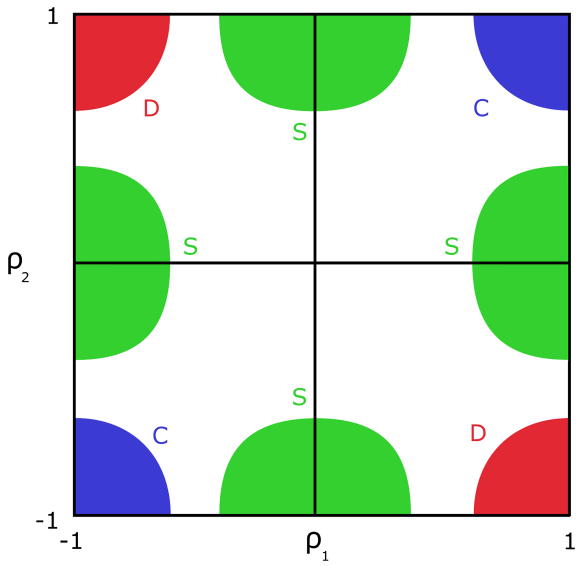


Figure 2.3: Score surface showing the combinations of correlation coefficients corresponding to three types of differential co-expression; C, S and D [3]. C (blue) is conserved (strong co-expression in both conditions with no sign change), S (green) is specific (strong co-expression in only one of the conditions), and D (red) is differentiated (strong, but oppositely signed co-expression values). ρ_1 and ρ_2 denote the Spearman's rank correlation of a given gene pair in condition 1 and 2, respectively. Only the values within the colored areas end up as links in the network. Image from Voigt et al. [3], under the CC BY 4.0 licence.

The final step is to transform the similarity matrix into an adjacency matrix, the fundamental element for network construction. To do this, one needs to decide a threshold for what is considered a significant enough change, and in this framework a "hard" thresholding-algorithm is performed. This means deciding a cut-off value τ by which all gene pairs with similarity values lower than this are rejected. However, since the three similarity scores C_{ij}, S_{ij} and D_{ij} have different distributions[3], the scores are not directly comparable. Therefore one single threshold value is not solid enough. Instead, three different cut-off values $X_p^{C,S,D}$ are calculated independently from each underlying distribution, based on a common sampling approach:
First, m samples s_{ij} of size $L \ll M$ is drawn from each pool of M total similarity scores C, S or D. From this, each threshold value $\tau = X_p$ is defined as the average of the maximum score per sample [3]:

$$X_p = \frac{1}{m} \sum_{i=1}^{m} \max_{\{s_i\}} X \tag{2.13}$$

To allow for meaningful comparisons in the final network, the three cut-off values are based on a common *importance level p*, which maps the similarity scores onto a common scale. This is set as $p = 1/L$, which means that it is only dependent on the chosen sample size L, which must be the same for all three similarity scores. The importance level can be adjusted to give a network size suitable for downstream analysis. This adjustment corresponds to changing the colored areas in Fig. 2.3, by increasing or decreasing the amount of total edges included in the network. For each gene pair, at most *one* of the similarity scores can be above their own threshold, ensuring that only one type of differential co-expression link ends up in the network, classifying the gene pair into the most appropriate category [3].

Even though a gene pair can only have *one* link type, a gene (node) in the network can be connected to multiple other nodes with C-, S- or D-type of interactions. In order to determine the fraction of the three different link types that a node has to its neighbors, we can use a score called *node homogeneity H*:

$$H_i = \sum_{j \in \{C,S.D\}} \left(\frac{k_{j,i}}{k_i} \right)^2, \tag{2.14}$$

where $k_{C,i}, k_{S,i}$ and $k_{D,i}$ is the number of C, S and D-type interactions that node i has, respectively, and k_i is the nodes degree. The highest score is H = 1, which means that all of the nodes links are of the same type. The lowest score is H = 1/3, which indicates an even distribution of C-, S- and D [3].

2.5 Protein-protein interactions (PPIs)

Interactions between proteins are fundamental for all cellular processes, including signal transduction, metabolic pathways, and cell cycle progression [10]. Therefore, investigating protein-protein interactions (PPIs) is crucial for comprehending biological

functions both in healthy and disease states. The complete map of all possible protein interactions that can occur in a cell or organism is called the *interactome* [64]. Although the human interactome is still far from complete, it allows for interrogation of how proteins and protein complexes work at a system-level [64]. This proteome-scale collection of PPIs can be represented as a complex network where nodes and undirected links constitute proteins and the interactions between them, respectively. The PPI network possesses many of the same characteristics as other biological networks [7], and can be studied using the same network parameters explained in the Network Theory (section 2.2). It is important to distinguish between an *interaction* - commonly understood as a direct physical contact - and other relationships/associations that indicate a shared function but not necessarily physical binding. While the former are experimentally determined (known) interactions, the latter are *predicted* interactions, for example from co-expresson data [65]. De Las Rivas et al. proposed the following definition of PPIs: "specific physical contacts between protein pairs that occur by selective molecular docking in a particular biological context" [65]. PPIs have varying stability; some are stable/permanent, while others are highly transient. Protein complexes, such as ATP synthase, often involve stable PPIs between the subunits working together as a cellular machine. Other processes can rely on the brief interactions between several cascading proteins, such as in signaling pathways. Another important aspect is the biological context. PPIs depend on factors such as cell type, environment, post-translational modifications, cofactors, and other binding partners [65].

The number of PPIs reported has increased considerably in recent years as a response to the development of efficient high-throughput techniques [65]. Many public databases are available that integrate PPIs from multiple studies, both small-scale and large-scale. There are many methods available for detecting PPIs, with the yeast two-hybrid (Y2H) system being one of the most renowned experimental techniques. Y2H is a binary screening method for detecting pairwise protein interactions [10]. It was first described in *Saccharomyces cerevisiae*, and the method utilizes the transcription activation system in yeast [37]. The two proteins tested (called "bait" and "prey") are fused separately to the two domains of the transcriptional activator Gal4. Only if bait and prey interact do they restore the function of Gal4, which results in gene expression of a chosen reporter gene. Hence, the PPI can be inferred by measuring the resulting products of the reporter gene expression. Often the *HIS3* reporter gene is used, which only when activated produces the necessary histidine amino acid for yeast growth [10]. The use of yeast as host system presents some limitations. First of all, it does not account for post-translational modifications that would happen in human cells, and therefore leads to false negatives. The method is also associated with a large rate of false positives. Even though proteins physically bind when "forced" together, they might never do so inside cells, due to different localizations or lack of simultaneous gene expression. Several variants are being introduced to handle these challenges, but this brief introduction is sufficient for this work [37, 10, 65].

Chapter 3

Materials and methods

This chapter provides the materials and methods used in this thesis. The first sections describe the AD transcriptomic data material and the subsequent pre-processing and sample selection. Then, a detailed description of the implementation of the CSD framework on the chosen gene expression profiles is provided. The resulting networks are visualized in *Cytoscape*, and node- and network parameters calculated. Further, the network is partitioned into modules, and hub-genes identified. Functional enrichment analyses extract biological insights for the investigation of disease association. Differential mean expression is calculated from the original microarray data, and the results integrated as a network layer on top of the differential co-expression network (CSD network). Finally, a high-quality PPI network was downloaded to look for relations between co-expressed genes in the network and their potential PPIs. An overview of the complete methodology is visualized in Fig. 3.1.

The method developments and calculations in this thesis have largely been made using a combination of the programming language Python (version 3.7.4) and terminal commands in the Ubuntu Linux distribution (version 18.04.2 LTS). The main scripts used can be found on Github [66].

3.1 The AD microarray data

The AD data used in this thesis, accession number E-GEOD-48350 was downloaded from ArrayExpress database [67]. ArrayExpress is a public database of microarray-based gene expression data at EMBL-EBI[1] [68]. The gene expression profiles were generated by single-color microarray as part of the experiment *Microarray analysis of Alzheimer's disease patients across 4 brain regions*. The data set has been used in several studies [69, 70] and was made public on 21 April 2014. The raw data can also be found in the

[1] European Bioinformatics Institute, part of the European Molecular Biology Laboratory

Gene Expression Omnibus (GEO) by National Center for Biotechnology (NCBI), with accession number GSE48350 [71].

The overall design of the microarray experiment performed by Berchtold et al. [67] was as explained in the following paragraph. Postmortem brain tissue was collected from human brain banks at the Alzheimer's & Dementia Resource Center (ADRC). In total, 80 samples from patients diagnosed with AD and 173 healthy controls were gathered. Detailed information of samples can be found here: ebi.ac.uk/arrayexpress/experiments/E-GEOD-48350/samples. From each sample, total RNA was isolated and transcriptome profiling performed by high-density oligonucleotide expression arrays. The microarray platform used was Affymetrix Genechip^{TM} Human Genome U133 Plus 2.0 (GPL570). The chip originally comprised of more than 54000 probe sets per sample, recognizing more transcripts than earlier microarray platforms. This was reduced to 21060 points by mapping to their respective genes in the platform. Normalization was performed by GC-RMA as described by Wu et al. [72], an algorithm similar to Robust Multi-Array analysis (RMA), but using a model based on GC (Guanosine-Cytosine)-content. Then the values were log-transformed with base 2, in order to make the data symmetrical for more accurate comparisons, especially important for differential expression analysis (see section 2.3.2). For more details, see the protocol of the experiment [67].

3.2 Data pre-processing and sample selection

The AD data material and sample information was accessed as described in detail above. It was chosen for this thesis based on the high total sample count (n = 253), including 80 samples from patients diagnosed with AD and 173 healthy controls. In order to make the data set ready for CSD implementation, it was quality-checked and further processed. The code written in Python, called preprocess.py, can be found on Github [66]. All probes mapping to more than one gene were removed to adjust for cross-hybridization. The resulting unique probe IDs were converted into gene symbols based on annotation according to platform GPL570. For genes with multiple associated probes, the mean of the intensity values of the duplicated probes was calculated. The resulting data matrix after filtration and gene-level summarization had 21044 unique genes (rows).

Samples were chosen from the data according to the aim of this thesis, namely to identify genes involved in AD. Hence, a single factor study design was used for all further analyses, and other potential confounding factors were intended minimized. The goal was to find genes that behave differently in individuals with AD, compared to those that do not have the disease. In the full data set, several factors other than disease could influence the analyses, and these were therefore considered carefully. Such experimental factors include age, sex and organism part. Age was considered the most important potential confounding factor, since it is a well-known risk factor of AD. Similar age span for both case and control was therefore ensured to minimize the effect of age-related gene expression. All AD patients were >60 years old (age span 60-95, mean 85.39 $\pm$ 7.37 years). Only age-matched controls were included (age span of 64-99 (mean age 81.27 $\pm$ 10.10 years), which reduced the control sample size from 173 to 93. Controlling for age

was estimated to have a large impact on the resulting differential co-expression network, and this effect was verified by performing the CSD method on the full data set before filtration for comparison (see appendix F). The sex ratio was 129/124 female/-male for the full data set, and 95/78 after filtering out the young controls, considered quite well-balanced. The postmortem tissue was extracted from four different regions of the brain - hippocampus (HC), entorhinal cortex (EC), superior frontal gyrus (SFG) and postcentral gyrus (PCG). For the generation of the differential co-expression network, these samples were pooled together in order to maintain a high sample size of n = 173 (80 AD, 93 control). The network would therefore only consider disease-related changes in brain tissue as a whole.

Based on the experimental design, the desired sample columns with their mean expression values were then extracted from the original data set into new text-files for downstream analysis. Each text-file contained either disease- or control-expression values, all values being $\log_2$ ratios. For implementation into the CSD framework (next section), it was important that these files had the correct format; a header and tab-separated columns. The first column included the gene symbol and the remaining columns contained the average expression of the transcripts for each of the 21044 unique genes in the different samples (one row for each gene).

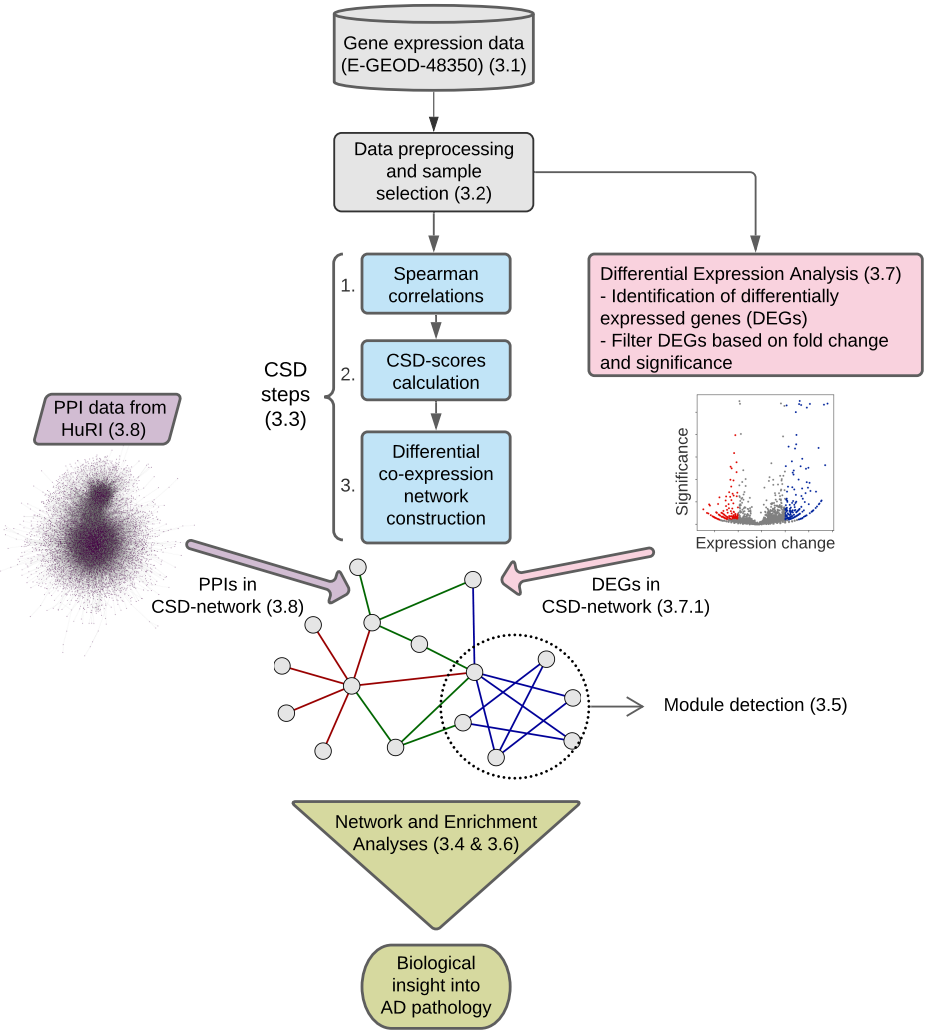


Figure 3.1: Overview of methodology. The flowchart shows the steps performed in this thesis from transcriptomic Alzheimer's disease (AD) data accessed in the expression database to the end goal of biological insight. The three steps of the CSD method for generating a differential co-expression network (example network made in Cytoscape) are shown in blue. The two main additional analyses integrated are represented in purple and pink. The section numbers explaining each process are shown in parentheses. PPIs: Protein-Protein Interactions. PPI network shown here (left) is the giant component of the HuRI (Human Reference Interactome) used in this work, visualized in Cytoscape. Volcano plot on the right side was modified from [73].

3.3 Differential co-expression network construction

The differential co-expression network analyzed in this thesis was constructed using the CSD framework developed by Voigt and Almaas [3]. It includes a set of three software programs, and the original software can be found at [74]. Adjustments made to the original framework will be described here in detail. The programs were run on a HUNT Cloud virtual machine [75], which had sufficient amount of memory storage and processing power for this large-scale analysis. Specifically, the almaaslab-compute3 machine with 16 CPU cores and 295 GB RAM was used [75]. The three scripts were run separately and performed each of the three main steps (blue in Fig. 3.1) in the differential co-expression analysis (explained in section 2.4.1):

1. Correlation coefficients calculation
2. CSD-scores calculation
3. Network generation

An important change to the original framework was that correlation variance calculation through sub-sampling was omitted from the implementation. It has been shown earlier that skipping the variance correction has little influence on the robustness of identifying disease-related genes for varying sample sizes compared to conventional CSD [17]. It was therefore considered reasonable to have no sub-sampling in order to decrease the running time of the program substantially. The scripts were changed accordingly, removing all variables and lines of code associated with sub-sampling calculation. All modified code is available at Github [66].

The first step in the CSD-implementation was the calculation of intra-cohort similarity, using the Spearman's rank as correlation coefficient. This was performed by the C++ script called FindCorr.cpp. Some parameters were changed depending on input file. The sample size was set to 80 for disease data and 93 for control data, corresponding to the number of data points per gene in each text file. The number of genes was set to 21044 for each input files. Once compiled, the script calculated the pairwise correlation for all gene combinations in each expression file individually - one time for the disease set and one time for controls. It was imperative that the two output files had matching gene pairs line-by-line for the next step to be successful. This was ensured by the original expression data input having the same number of genes sorted in equal order.

The output text-files from the two iterations of FindCorr.cpp were then used as input in FindCSD.py. This Python script compared correlation values from the two conditions and calculated the C-, S- and D-scores using the equations (2.10)-(2.12). The variance variables were removed from the script, and the denominators of the similarity scores (comboSD in the script) were set to 1. This restricted the range of the C,S,D-scores to [0, 2] for C_{ij} and [0, 1] for S_{ij} and D_{ij}. The script produced a file containing correlation under both conditions, as well as gene relationship scores for all gene pairs. It also produced three files including only C-, S- or D-values for each gene pair, that were used for network generation in the next step.

In the final step of the CSD framework the Python script CreateNetwork.py was used

to generate four networks; one for each interaction type and one aggregate network (the latter is exemplified in Fig. 3.1). Only the similarity scores above their respective cut-off values, calculated with eq. (2.13), were included as interactions (edges) in the network. The stringency of this cut-off was determined by the parameter selSize in the code, corresponding to the sample size L used for determining a common *importance level p*. Several different selSizes were tested in order to obtain a network with a size appropriate for further analysis. The one used was $L = 2 \cdot 10^5$, yielding an importance level of $p = 5 \cdot 10^{-6}$. The output files had the same format: each line representing an edge in the network, with the co-expressed genes in the first two columns and their similarity score value and interaction type (C, S or D) in the third and fourth column, respectively. The text-files were imported into *Cytoscape* for visualization.

3.4 Network analysis

Network topology was investigated using the software environment *Cytoscape v3.7.2* [76]. Genes were represented as circular, gray nodes unless stated otherwise, and labeled when appropriate. The edges in the network were color-coded based on the type of differential co-expression: C=blue, S=green and D=red. The default layout (Prefuse Force Directed) was used unless stated otherwise. The included Cytoscape tool called *NetworkAnalyzer* was used to calculate several node- and network parameters, such as node degree, degree distribution and assortativity. The values for degree distribution and neighborhood connectivity distributions were transferred from Cytoscape and read into Python for plotting with matplotlib [66].

3.4.1 Node homogeneity

The self-written Python script *homogeneity.py* was used to calculate the fraction of the three different link types that each node has to its neighbors. This node homogeneity (H) was calculated using Eq. (2.14) on all genes in the CSD network. The homogeneity was then plotted as a function of node degree, represented by a boxplot. The boxplot was generated using the function "boxplot" in the matplotlib.pyplot interface [77]. Default parameters were kept, except for the boolean value "showmeans" which was set to "True" to show the arithmetic mean values of homogeneity per degree as green triangles. The relative number of genes involved in each interaction (co-expression) type in the network was presented in a Venn diagram. These numbers were calculated by comparing the list of genes in each of the three individual networks using variations of the "grep -x -f file1 file2 | wc -l" command in Ubuntu terminal.

3.5 Module detection

For further detailed study into the underlying biology of the network, it was of interest to reduce the global network complexity by focusing on more manageable sub-graphs of the network. The complete CSD network was therefore partitioned into smaller subunits of highly connected nodes, so-called modules (see example in Fig. 3.1). The Lou-

vain community detection algorithm [45] was used to identify modules in the CSD network. To implement this in Python, the *Python_louvain*-module was first installed. It is dependent on NetworkX [44], which was used to read the complete network as an edgelist. The network was partitioned by using the "best_partition"-function in the *community_louvain* package. This function returns an index for each gene, representing their module affiliation. As explained in section 2.2.6, the Louvain algorithm detects modules by optimizing modularity. To evaluate the quality of the partition, the "modularity"-function was used to calculate the global modularity score, Q, based on Eq. (2.6). The module indices were then imported into Cytoscape as node attributes in the CSD network for further analysis.

3.6 Functional Annotation and Enrichment Analyses

For the most prominent genes in the network, the GeneCards Human Gene database [78] was used to identify the annotated biological function of these genes individually. Even though all genes in the network would essentially be interesting to examine as they have passed the thresholds for co-expression, the biological interpretation of thousands of genes is still a challenging task. Therefore, several bioinformatics tools with biological knowledge from public databases were used to systematically find the most enriched and relevant biology from the network.

Functional enrichment analyses of gene sets (both within networks and modules) were performed using the Gene Ontology Consortium tools [79, 80]. This is a web-based service that performs over-representation tests powered by the PANTHER classification system [81]. A list of genes is taken as input and the test looks for genes with biological associations that are either over- or under-represented in the gene set compared to what would be expected by chance. A background reference list, which by default is all the genes from Homo Sapiens in the PANTHER database, was used for comparison. The ontology term "Biological Process" was selected for the search. Genes found to be enriched for a biological process would indicate that their gene products have molecular functions involved in that certain biological program [82]. It is a diverse concept and the GO terms can be as broad as "signaling" or more specific like "negative regulation of synaptic transmission". If not stated otherwise, only the most specific terms are presented. To search for over-represented *Pathways*, the 2019 Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway library in Enrichr was used [51, 83]. KEGG has a widely used collection of pathway maps for analyzing systems of interconnected molecular reactions and interactions [4].

The enrichment analysis returns a list of all the significant terms including the results of statistical testing. The Fisher's exact test with FDR correction, as calculated by the Benjamini-Hochberg method, was used to determine q-values. A threshold of FDR $<$ 0.05 was used to determine if the mapping of genes to a certain GO term or KEGG pathway was statistically significant. The output was sorted by Fold Enrichment (FE) of the most specific categories. The FE score represents the extent to which the number of genes in the provided list is greater (+) or lower (-) than the expected number of genes involved in the annotated biological process or pathway [79, 80].

3.6.1 Disease Association

There are a lot of studies revealing almost endless genes associated with AD. Many different databases combine experimental data with curated articles. In this thesis the MalaCards database [25] was chosen for the search, as it is based on the GeneCards database which was used for individual gene annotation. It is an integrative database which combines knowledge from a substantial amount of sources, including the commonly used databases Online Mendelian Inheritance of Man (OMIM) and Uniprot [25]. As stated in the MalaCards website, a gene is identified as associated with AD based on i) the GeneCards search mechanism, ii) genetic testing resources supplying specific genetic tests for the disease, iii) genetic variations resources supplying specific causative variations in genes for the disease and iv) resources that manually curate the association of the disease with genes [25]. For details on the annotation schemes, the reader is referred to the original article [25]. The list of 499 AD-related genes was compared for overlap in the original microarray data and in the constructed differential co-expression network. Fold enrichment and p-values were calculated using the cumulative distribution function (CDF) of the hypergeometric distribution. AD-related genes in the network were highlighted as diamond-shaped nodes.

3.7 Differential expression analysis

The next step in the transcriptomic analysis was to look for differentially expressed genes (DEGs) in the microarray dataset (pink processes in Fig. 3.1). This was to identify genes that were significantly up- or down-regulated when comparing AD patients to control, as these could be essential for disease progression. DEGs were identified from the normalized $\log_2$-transformed gene expression values using fold change (FC) and multiple testing. The Python script for performing these calculations on chosen input data sets is given in appendix D. First, the mean expression level of each gene across all samples was calculated for case and control files, separately. Because the data was already $\log_2$-transformed, the FC was then calculated by taking the difference between the two means directly. To test whether each mean level differed significantly across samples, independent t-tests were performed using the ttest_ind from the scipy-library in Python. This built-in function returned the T-statistic and the raw p-values, indicating the significance of individual tests. To adjust for multiple comparisons, the Benjamini-Hochberg (BH) method was used to control the false discovery rate, calculating the adjusted p-values (q-values) with Eq. (2.8). Genes were considered DEGs when the q-value was < 0.05 and $|\log_2 FC| > \pm 0.2$.

To visualize both the magnitude and significance of change in gene expressions, a volcano plot[2] was generated where the negative log of the q-values for all genes between healthy and AD-individuals was plotted against the $\log_2$FC. This plot was chosen because it can visualize the expression of thousands of genes (here 21044 genes) at the same time, while highlighting significant DEGs. The plot was generated in Python using *Bioinfokit*, a bioinformatics data analysis and visualization toolkit [73]. Significantly

[2] 2D scatter plot with shape like a volcano

up-regulated genes were colored red and down-regulated genes blue, as this is increasingly common to do to avoid the color-blindness issue that arises from the traditional red-green color scheme.

3.7.1 DEGs in the CSD network

It was further of interest to investigate the overlap between DEGs and differentially co-expressed genes in the CSD network. To inspect the location of DEGs in the CSD network, the csv-file with all values from the DEA was uploaded as a node attribute table to the CSD network in Cytoscape. The size of the nodes in the network was sorted by $\log_2$FC, using continuous mapping, but a minimal node size was set for all nodes that were not DEGs. Nodes with significant differential expression (FDR < 0.05) were also colored continuously based on the sign of $\log_2$FC - from decreased (blue) to increased (red) expression. The greater the magnitude of $\log_2$FC, the darker the color. The nodes with too low magnitude of change ($|\log_2 FC| < \pm 0.2$) were kept gray-colored.

3.8 Integration of Protein-Protein Interactions

To integrate protein interactions as a network layer with the CSD network (purple process in Fig 3.1), a map of human reference protein interactions was collected from the Center for Cancer System Biology's (CCSB) database. Their newest collection of high-quality PPI data, and the largest of its kind to date, was used: HI-Union [84]. This dataset is an aggregate of all PPIs identified from several CCSB mapping efforts: HI-I-05, HI-II-14, HuRI, Venkatesan-09, Yu-11, Yang-16, and Test space screens-19 [84]. All contain high-quality binary interactions generated through systematic mapping of open reading frames (ORFs) by yeast two-hybrid (Y2H)-assay [84].

The PPI data was downloaded as a tab-separated file with the interacting proteins being indicated as pairs of Ensembl gene IDs. In order to compare the PPIs with the gene pairs of the CSD network, the CSD interactions were converted from gene symbols to Ensembl IDs. The two edge lists were imported to Cytoscape and the networks merged to look for overlap. The search for edges present in both CSD and PPI was also done by using "grep -x -f" in Ubuntu terminal, checking for common lines in the two edge-list files. For easier interpretation, the IDs were converted back to gene symbols. The network on the left in Fig. 3.1 shows the giant component of the PPI network used.

Chapter 4

Results and analysis

This chapter presents the differential co-expression network generated from the transcriptomic AD data and the biological analyses performed on it. It is divided into two parts; section 4.1 provides the analyses performed directly on the CSD network, and section 4.2 includes the integrated analyses; DEA (4.2.1) and PPIs (4.2.2).

4.1 CSD framework on AD Expression Data

This first part starts with a description of the overall topology of the network and evaluates network parameters such as homogeneity, degree distribution and assortativity. Functional enrichment is considered both on network level and after modular decomposition. Finally, the analysis was narrowed down to a study of the most prominent genes in the network and their potential disease association.

4.1.1 CSD network construction and visualization

The CSD framework was implemented on gene expression data from brain tissue in AD patients ($n = 80$) vs healthy controls ($n = 93$) to generate a differential co-expression network. From the total of 21044 expressed genes in the data set, giving rise to more than 200 million ($\frac{n^2}{2}$) co-expression link combinations, 2044 links were extracted as significant based on the importance level of $p = 5 \cdot 10^{-6}$. Fig. 4.1 visualizes the resulting aggregated network, with 1535 nodes (genes) and 2044 links, of which there is an evenly distributed number of C-type (709), S-type (690) and D-type (645). The networks with individual C-, S- and D- type links are visualized in Fig. 6.1-6.3 in appendix A. While it is important to note that each figure shows only one of many possible Cytoscape visualizations of the same network, the connections and network properties remain the same for each visualization.

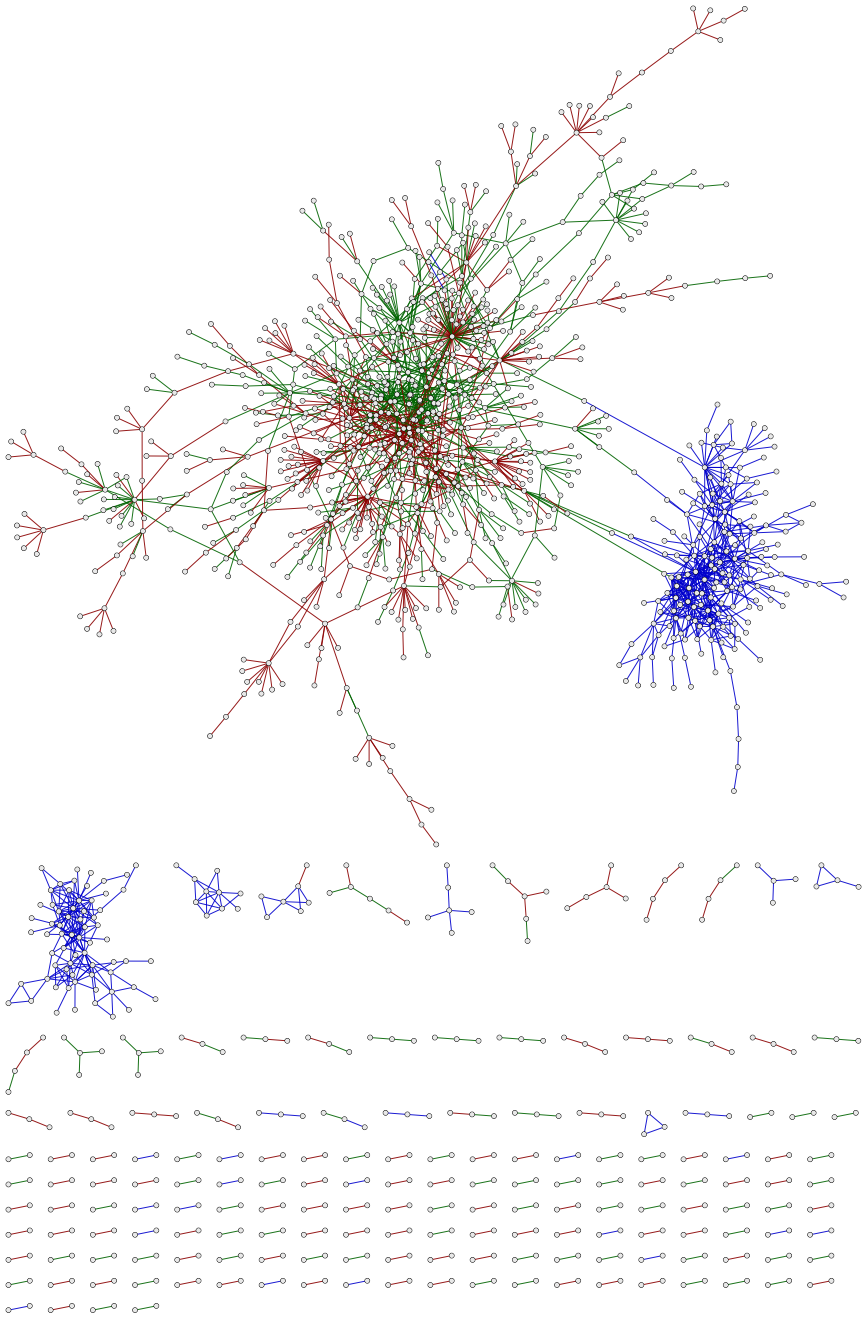


Figure 4.1: CSD network. Visualization of the aggregate differential co-expression network from transcriptomic data (80 AD patients, 93 controls). $N = 1535, M = 2044$. Nodes represent genes and links represent the type of co-expression between pairs of genes. Links are colored by type: blue is conserved (C), green is specific (S) and red is differentiated (D). Network generated using an importance level of $p = 5 \cdot 10^{-6}$ and visualized in *Cytoscape*.

The first striking feature observable from the CSD network (Fig. 4.1) is that the conserved links (blue) are generally separated from the specific and differentiated links, which are highly interconnected. The giant component of the network, containing 1078 nodes and 1636 links, is split in two regions. The largest of those contain an even mix of S- (608) and D-type (549), while the smallest region has almost exclusively conserved links and is connected to the largest region only by a few "bridge" nodes. The two next largest connected components are both entirely C-linked, with 66 and 10 nodes respectively. The majority of the remaining 161 connected components form 127 isolated pairs. Although the visualization of the network is a valuable starting point, it could be misleading to solely focus on the visual aspects. Further analyses of network properties and central nodes are essential to develop a deeper understanding of the potential underlying biological processes that are not visible by merely looking at the network.

4.1.2 Node homogeneity

The node homogeneity was calculated for each node individually in the CSD network, and the results are summarized in Fig. 4.2. These results show that there is an overall tendency of links with the same co-expression type to group together, in other words the overall node homogeneity is high. From the Venn diagram in Fig. 4.2(left) we can see that no node in the complete network is fully heterogeneous with links of all types. This is also evident from the box plot (Fig. 4.2(right)) which shows that the lowest value is 0.5 and not 1/3 which would indicate full heterogeneity. From the box plot we see a trend of mean homogeneity values (green triangles) of $H \geq 0.8$, independent of degree. However, there are two outliers to this trend, which correspond to two of the largest hubs, with degrees $k = 69$ and $k = 31$, they have relatively low H-values, $H = 0.54$ and $H = 0.65$ respectively. Of all nodes with mixed interactions, the combination of specific and differentiated is the most common (N=163). This is reasonable as these two type of links are both differential, caused by a change of co-expression pattern from control to case. The conserved (blue) links are highly interconnected and separate from the other two types of differential co-expression, as seen clearly from Fig. 4.1, and summarized by the numbers in the Venn diagram (Fig. 4.2(left)). Only 9 nodes with C-links are also connected to nodes with either S- or D-type links.

4.1.3 Degree Distribution and Degree Correlations

The degree distribution and degree-degree correlations are both essential measures for characterizing overall network topology based on the degree k. These parameters were mostly used to verify earlier findings on co-expression network structure.

Degree Distribution

The node degree distribution of the total differential co-expression network is shown in Fig. 4.3. The number of nodes plotted as a function of degree follows a power law, as approximated by the straight (red) line in the log-log plot with a slope of $-1,876$ and an R^2-value of 0.895. This suggests that the network, as expected from biological

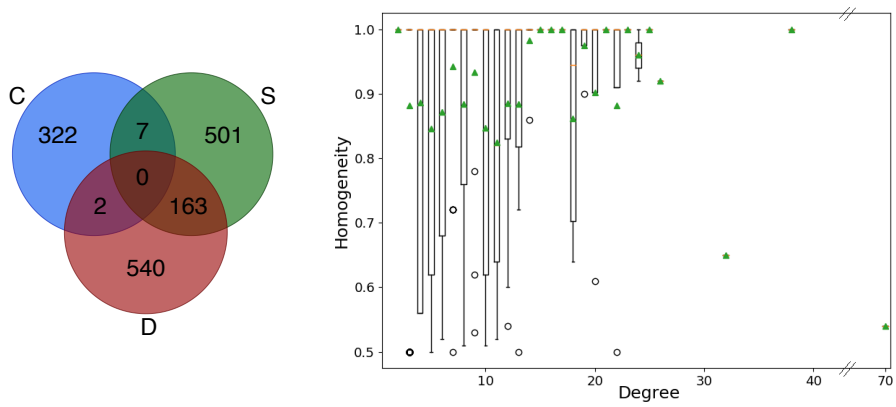


Figure 4.2: Node homogeneity. **Left:** Venn diagram of the relative number of genes involved in each type of interaction (co-expression). Blue = conserved (C), green = specific (S) and red = differentiated (D). **Right:** Box plot of node homogeneity binned by node degree. The boxes go from the first quartile (25th percentile) on the bottom to the third quartile (75th percentile) on top. Median values are represented by red bars and mean values by green triangles. The ends of the whiskers correspond to the minimum and maximum values of H for the given degree.

networks, has a scale-free topology. Although the average degree of the network is $< k >= 2.66$, no "typical" node exists, as compared to random networks where most nodes have the average degree. In scale-free networks most nodes have low degrees, but at the same time the probability of observing high-degree nodes is substantially higher than random. This suggests that the correlations between the genes in the co-expression are not random but has biological function.

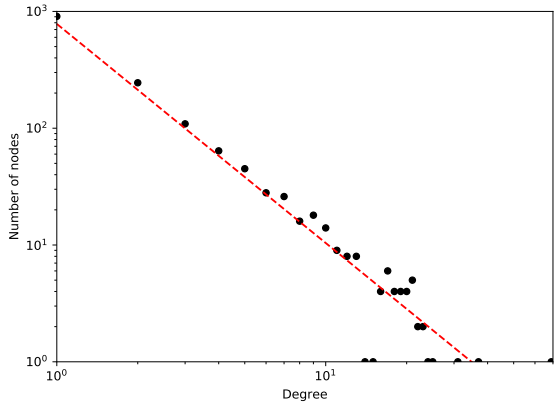
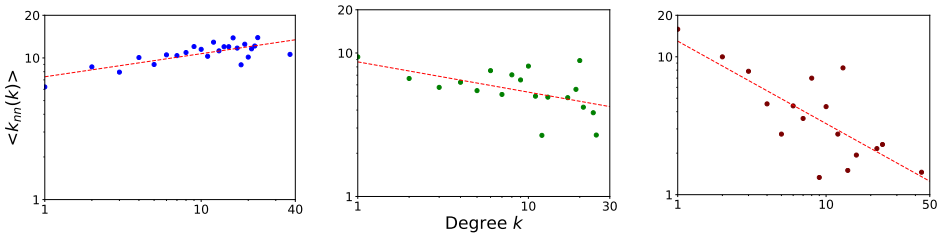


Figure 4.3: Degree distribution of the CSD network. The number of nodes as a function of degree on a log-log scale. A power law of the form $y = 782x^{-1.876}$ was fitted with $R^2 = 0.895$ (dotted red line).



(a) C-network. **(b)** S-network **(c)** D-network

Figure 4.4: Neighborhood connectivity distributions for the individual C-, S- and D-networks. The average degree of nearest neighbors of a node (<$k_{nn}(k)$>) as a function of node degree (k) on log-log scale. Red dotted lines are power laws fitted to the data points: **a)** $y = 7.36x^{0.163}$ (correlation: 0.706 ($R^2 = 0.580$)), **b)** $y = 8.67x^{-0.211}$ (correlation 0.539 ($R^2 = 0.288$)),**c)** $y = 12.99x^{-0.598}$ (correlation: 0.869 ($R^2 = 0.594$)).

Degree Correlations

To qualitatively determine how the nodes connect to other nodes with a certain degree, the neighborhood connectivity distribution $k_{nn}(k)$ was plotted for each individual C-, S- and D-network (Fig. 4.4a-4.4c, respectively). From the log-log plots an approximated power law fit returned the degree correlation exponent μ, which tells us if the network has an assortative, neutral or disassortative nature. The assortative mixing of the CSD network as a whole was not considered, as the three individual types of co-expression have previously shown to have quite different characteristics of degree correlation [3]. As expected, the C-network is slightly assortative (Fig. 4.4a), as seen from the increasing degree correlation function ($\mu = 0.163$). This means that the nodes have a tendency of connecting to other nodes with a similar number of neighbors as themselves. On the other hand, both the specific and differentiated network show disassortative topology (Fig. 4.4b,c), but the D-network to a much higher extent. The neighborhood connectivity distributions of the S- and D-networks are decreasing, with negative correlation exponents of $\mu = -0.211$ and $\mu = -0.598$, respectively. Negative degree correlations indicate that there is a hub-and-spoke topology, where hubs tend to connect to low-degree nodes and form "bouquet"-like structures. This is especially evident in the sub-network with exclusively differentiated links, which can be revealed from Fig. 4.1, but even more clearly from Fig. 6.3 in Appendix A.

Hubs

As the degree distribution shows, a few genes in the total network have considerably higher degrees than average. These *hubs* are interesting to investigate as they contain a major part of the interactions in the network, and contribute to most of the topology. A limit of $k \geq 20$ was used to define a node as a *hub* in the CSD network, and this resulted in 18 hubs. These represent only 1.17% of the nodes in the network, but contain as much as 21.5 % of the links. Table 4.1 displays these hubs sorted by degree (k), along with the number of each link type (k_C, k_S and k_D) and their calculated node

Table 4.1: Network hubs. Genes in the CSD network with node degree $k \geq 20$. $k_{C,S,D}$: degree of interaction type C,S,D. H: Node Homogeneity.

Gene	k	k_C	k_S	k_D	H
KIAA1841	69	0	25	44	0.54
NMNAT2	37	37	0	0	1.0
MIGA2	31	0	24	7	0.65
AQR	25	0	1	24	0.92
AL158206.1	24	0	0	24	1.0
HPRT1	23	23	0	0	1.0
GTF2I	23	0	1	22	0.92
TOM1L2	22	0	0	22	1.0
YWHAH	22	22	0	0	1.0
GOT1	21	21	0	0	1.0
NAPB	21	21	0	0	1.0
TMEM178A	21	0	21	0	1.0
PLTP	21	0	20	1	0.91
LCAT	21	0	11	10	0.5
ENPP2	20	20	0	0	1.0
CADPS	20	20	0	0	1.0
MDH1	20	20	0	0	1.0
VSNL1	20	20	0	0	1.0

homogeneity score (H). As we can see from the table, most of the hubs are homogeneous (H close or equal to 1.0), connecting to most or all of their neighbors with the same co-expression type. The nine hubs with conserved (C) type of interaction to their neighbors, meaning that they have strong co-expression under both conditions, are all homogeneous. Hence all the heterogeneous hubs with $H < 1.0$ have a mixture of S- and D-type of links. This is not surprising as the conserved regions (blue) are separated from the differential regions (red and green), as we saw earlier from Fig. 4.1. Five hubs, namely KIAA1841, AQR, AL158206.1, GTF2I and TOM1L2, are predominantly linked to their neighbors by differentiated (D)-type. This indicates that these genes have strongly, but oppositely signed co-expression under the two conditions. For the remaining four nodes, representing the genes *MIGA2*, *TMEM178A*, *PLTP* and *LCAT*, the majority of the interaction types are specific (S). This means that the co-expression is present only under one condition. By looking more detailed into the calculated Spearman correlations, it was found that all of these S-interactions go from no/weak correlation in control to strong correlations in sick patients. This might be due to a coordinated regulation of the genes in the disease. Before going into the biological importance of the individual hubs (see section 4.1.7), we will start with a more global analysis of functional processes on a network level.

4.1.4 GO Enrichment Analysis on C-, S- and D-networks

To give an overview of biological processes that each of the different types of co-expressed genes might be involved in, a GO functional enrichment analysis was performed on each of the individual subnetworks.

GO on the C-network ($N = 331$) gave the largest number of significantly enriched biological processes. Genes in this network have a conserved co-expression, meaning that the pairwise correlation patterns are unchanged for AD patients compared to healthy individuals. The enrichment therefore highlight processes that might be conserved between conditions and thus prominent for brain tissue or cell functions in general. Here we find genes enriched in processes such as nervous system development (FE = 2.70, $q = 2.17 \cdot 10^{-17}$), synaptic transmission (FE = 4.50, $q = 1.25 \cdot 10^{-8}$) and vesicle-mediated transport in synapse (FE = 10.13, $q = 1.21 \cdot 10^{-11}$). The complete table of results can be found in the following doi: 10.6084/m9.figshare.13342217 [26]. To further limit the overwhelming amount of information, a more local GO was performed on the smaller C-modules (see section 4.1.6).

The genes in the S- and D-networks have changes in pairwise correlation that indicate a disease-related change in the transcriptional program when comparing AD to control. No statistically significant GO terms were found for the list of genes in the D-network ($N = 705$). In the S-network ($N = 671$), some significant terms were found, but with substantially lower enrichment scores than for the C-network. The GO terms were also more general, mostly related to localization and signaling, and involving many genes. The three most specific terms were plasma membrane bounded cell projection organization (FE = 1.72, $q = 2.99 \cdot 10^{-2}$), phosphate-containing compound metabolic process (FE = 1.60, $q = 2.86 \cdot 10^{-3}$) and amide transport (FE = 1.60, $q = 3.27 \cdot 10^{-2}$). As these processes are quite broad, further investigation is needed to provide new information of mechanisms related to AD. The complete table of enrichment results is given in appendix B.1.

4.1.5 KEGG Pathway Enrichment

The 2019 KEGG Pathway database in Enrichr [51, 83] was used to search for overrepresented pathways in the network. The combined CSD network showed significant enrichment for 6 categories, of which 4 are related to signal transduction and the other two pathways are synaptic vesicle cycle and mineral absorption (Table 4.2). The signaling pathways enriched are major pathways central to many biological processes, important for the normal functioning of cells in general, including brain cells. It is therefore reassuring to find such pathways enriched, as these are important for the regulation of essential processes in brain tissue, where the analyzed samples were originally taken from. The broadest category, with the largest number of genes enriched in the CSD network, was PI3K-Akt signaling (FE = 1.69, $q = 0.0191$). The Phosphatidylinositol 3'-kinase(PI3K)-Akt signaling pathway regulates fundamental cellular functions such as transcription, translation, proliferation, growth, and survival [4]. The downstream effects of the phosphorylation cascade can be cell cycle progression or apoptosis, protein synthesis or glycolysis/gluconeogenesis [4]. MAPK- and insulin signaling are highly re-

Table 4.2: All KEGG Pathways significantly enriched in the CSD network, sorted by fold enrichment (FE). Overlap: ratio of genes found vs expected from the reference list. FDR: Benjamini-Hochberg adjusted p-value. Enriched genes are shown explicitly.

Term	Overlap	FE	FDR	Genes
Mineral absorption	11/51	2.81	0.0489	ATP1A2, ATP1B1, HEPH, HMOX1, MT1E, MT1F, MT1G, MT1H, MT1X, MT2A, TF
Synaptic vesicle cycle	16/78	2.67	0.0248	AP2M1, ATP6V0A1, ATP6V1A, ATP6V1B2, ATP6V1E1, ATP6V1F, ATP6V1G2, CLTC, CPLX1, DNM1, NSF, SNAP25, STX1B, STXBP1, SYT1, UNC13C
Glucagon signaling pathway	19/103	2.40	0.0220	ACACA, ATF2, CALM3, CPT1C, CREB1, CREB5, G6PC3, PCK1, PDHA1, PFKL, PGAM2, PHKG1, PPP3CB, PPP3R1, PPP4C, PRKAA1, PRMT1, PYGL, PYGM
Insulin signaling pathway	25/137	2.38	0.0126	ACACA, BRAF, CALM3, CBL, EXOC7, G6PC3, GSK3B, HK1, IKBKB, INPPL1, IRS1, MAP2K1, MAPK10, MAPK9, PCK1, PHKG1, PRKAA1, PRKAR1B, PTPRF, PYGL, PYGM, RAF1, RPS6, SHC1, TSC2
MAPK signaling pathway	42/295	1.86	0.0116	ANGPT1, ATF2, BRAF, CACNA1I, CACNB2, CACNG2, CDC42, CSF1R, DUSP3, ECSIT, ERBB2, ERBB3, FGF1, FGF9, FGFR3, FLNA, GADD45B, GNG12, HSPA2, IKBKB, IL1R1, JUND, MAP2K1, MAP2K6, MAP3K1, MAP4K4, MAPK10, MAPK7, MAPK9, MEF2C, MYD88, NF1, PAK1, PDGFD, PGF, PPP3CB, PPP3R1, RAF1, RASGRP3, RRAS2, TAB2, VEGFA
PI3K-Akt signaling pathway	46/354	1.69	0.0191	ANGPT1, ATF2, CCNE1, CCNE2, COL4A2, COL4A5, COL4A6, CREB1, CREB5, CSF1R, ERBB2, ERBB3, FGF1, FGF9, FGFR3, G6PC3, GNB2, GNG12, GNG3, GSK3B, IKBKB, IRS1, ITGB4, ITGB8, LAMA1, LAMA5, LPAR1, LPAR5, MAGI1, MAP2K1, OSMR, PCK1, PDGFD, PGF, PRKAA1, RAF1, RPS6, THBS3, THBS4, TLR2, TNC, TSC2, VEGFA, YWHAG, YWHAH, YWHAZ

lated pathways, which explains why there are several genes commonly enriched (Table 4.2). Genes might be involved in the regulation of different processes, leading to a great complexity in cell signaling. Alterations in these enriched signaling pathways that regulate the cell cycle can eventually lead to neuronal death, a common pathological feature in neurodegenerative diseases [5].

When searching for enrichment in the individual C-, S- and D-networks, no significant results were found in the S-only or D-only networks ($q = 0.5$). However, a long list of 44 significant pathway categories were found for the gene set of the C-network, which can be found here: 10.6084/m9.figshare.13342247. This shows that most of the enriched pathways in the combined network are caused by the genes in the C-network. It is not that surprising that these general pathways are potentially conserved, since these are vital for universal cell function. A dysregulation in these processes could lead to the progression of disease. The mineral absorption pathway had the largest fold enrichment in the CSD network (FE = 2.81, $q = 0.0489$). Six of the enriched genes were metalloth-

ioneins (MTs), proteins involved in metal homeostasis and oxidative stress response [85]. All are found in the C-network (Table 4.2, located in the small 7-node module third from the left in Fig. 4.1). This indicates that the genes form part of a protein complex which maintains metal homeostasis across conditions. The synaptic vesicle cycle was also largely enriched in genes from the C-network (FE = 10.8, $q = 1.07 \cdot 10^{-8}$, data not shown). In fact, 14 of the 16 genes shown in Table 4.2 were found in the C-network. Insulin- and MAPK signaling were also enriched in the C-network ($q = 0.009$). Although these two processes were not significantly enriched in the individual S- and D-networks, most of the genes enriched in these pathways in the combined network belong to the S/D-region of the network.

Interestingly, the glucagon signaling pathway was not significantly enriched in the C-network (3/103 expected genes, $q = 0.44$), but in the combined network it was (19/103, $q = 0.022$, Table 4.2). An overlap of 16/103 was found for the combined S- and D-genes (FE = 2.56, $q = 0.14$). This indicates that the genes enriched in glucagon signaling are mostly differentially co-expressed (S-type and D-type) in the network, and that the pathway might have an important role in the disease. Included in the list of enriched genes (see Table 4.2) we find *PHKG1* to have a prominent position in the network (indigo hub between node 1 and 12 in Fig. 4.6). It is found in the largest module, but has specific co-expression with two hubs of other modules, namely KIAA1841 and TMEM178A. It may therefore be interesting to look further into the role of this gene.

4.1.6 Module Analysis

In the hope of finding functionally related modular structures of co-expressed genes in the network, the Louvain community detection algorithm was applied. It resulted in 182 modules in total, however most of them with negligible sizes. Only the modules with 50 or more nodes were considered further. The resulting 11 modules were visualized in Cytoscape and presented in Fig. 4.5, where each module has differently colored nodes. The size, average degree, clustering coefficient and number of each co-expression type is summarized for each module in Table 4.3. This table also shows the largest hub of each module, but the biological function of these will be considered in section 4.1.7, as mentioned earlier. The global modularity score was as high as $Q = 0.83$, indicating an optimal partitioning of the network.

The community detection algorithm partitioned the giant component into 10 different modules, and the last module (turquoise) was the next largest connected component with 66 nodes (Fig. 4.5). The large S- and D-type region of the giant component was split into 8 separate modules. We see from the figure (Fig. 4.5) that these modules are quite sparse and spread across large areas of the region. As expected, the modules dominated by S- or D-type of interactions have clustering coefficients close to or equal to zero (Table 4.3). Oppositely, we find the three modules with different shades of blue, which include almost all the C-type links, to be more dense and clustered, forming stronger communities. These modules have clustering coefficients substantially higher than the average of the whole network ($C = 0.061$).

Module 1 (indigo) has as much as 150 nodes assigned to it, which is considerably higher

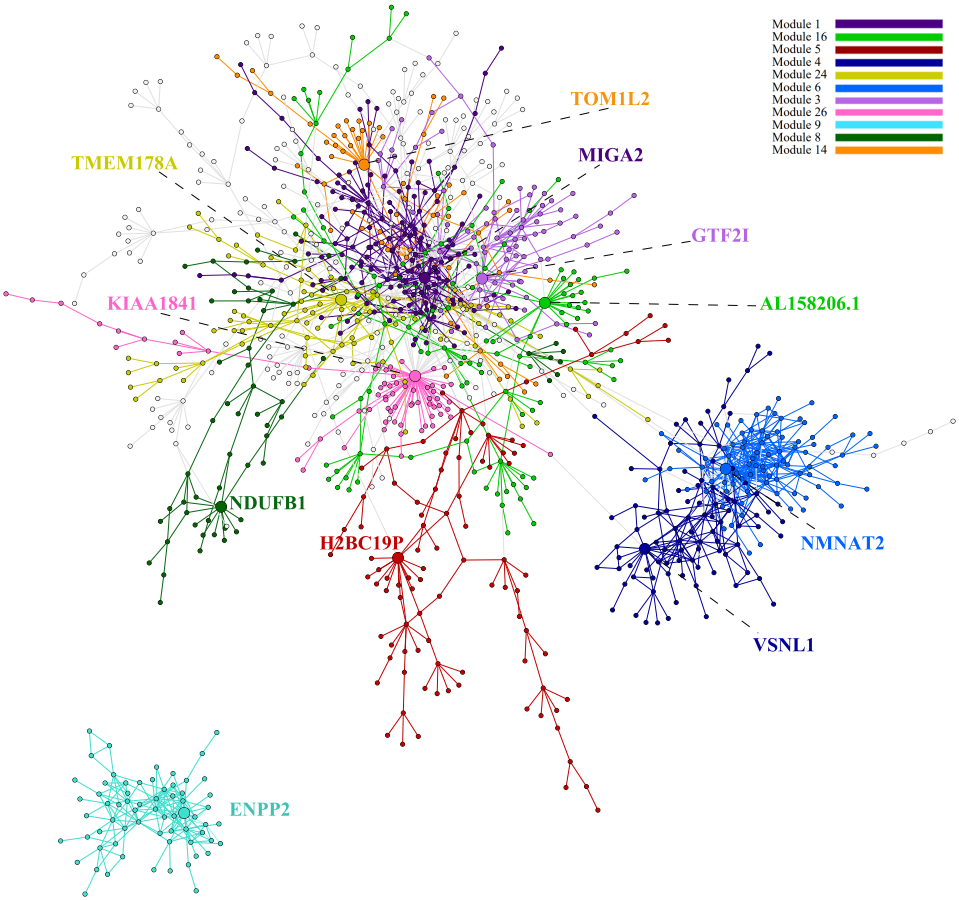


Figure 4.5: The 11 modules with 50 or more nodes, detected by Louvain algorithm, highlighted with unique colors in the CSD network. The color chart on the right side shows the assigned color to each module and their ID, sorted by module size (largest on top). The node with highest degree in each module is enlarged and color-labeled.

than the size of all the other modules. It is found in the center of the network in Fig. 4.5. The large majority of links in this module are of specific type, meaning that the co-expression between gene pairs is condition-dependent. The most highly connected gene in this module is *MIGA2* ($k = 20$), the third largest hub in the network. We also find the previously identified network hub PLTP in this module. PLTP has reduced importance in degree centrality ($k = 14$), but has the largest closeness centrality in the module. CYS1 was identified as a new node with modular importance, due to next largest degree ($k = 17$) and highest betweenness centrality. The values of CC and BC are not shown, as they are only relative and not directly comparable between modules.

Module 16 (green) is the next largest module, with 108 nodes. Here, there is a mixture of D-type and S-type of interactions, with the differentiated co-expressions dominated. The topology of the module is quite sparse, spreading over several of the other modules. The intra-modular hub of module 16 is AL158206.1, but another central node is NF2, which has both the highest closeness- and betweenness centrality in the module. It encodes Neurofibromin 2, a cytoskeletal protein involved in suppression of cell proliferation and tumorigenesis [78].

As we can discern from the visualization of the modular partition, module 5 (red) and module 8 (dark green) are the ones that look the most separated from the others in the S- and D- rich region (Fig. 4.5). Interestingly, we also identified new high-degree nodes from these two modules. Module 5 (red) has 95 nodes and about an even mix of S- and D-type of links (Table 4.3). The nodes form several dispersed, bouquet-like structures. The largest node by degree, forming an intra-modular hub is H2BC19P. It also has the highest closeness and betweenness centrality. *H2BC19P* is a pseudogene coding for a non-functional H2B Clustered Histone 19, a nucleosome component [78]. Although pseudogenes encode non-functional products, the processed transcript could be im-

Table 4.3: Module parameters. Modules in the CSD networks (ID colored as in Fig. 4.5) detected by Louvain algorithm with their number of genes (sorted by this), average degree (<k>), average clustering coefficient (C) and number of each link type ($k_{C,S,D}$). The largest hub of each module with its degree in the module is also presented.

Module ID	#genes	$< k >$	C	k_C	k_S	k_D	Hub
1	150	2.933	0.0	0	174	46	MIGA2 (k=20)
16	108	2.130	0.0	0	41	74	AL158206.1 (k=22)
5	95	2.042	0.011	0	44	53	H2BC19P (k=16)
4	92	3.630	0.211	167	0	0	VSNL1 (k=19)
24	90	2.289	0.0	0	77	26	TMEM178A (k=15)
6	86	6.116	0.383	263	0	0	NMNAT2 (k=28)
3	83	2.193	0.0	0	21	70	GTF2I (k=14)
26	68	1.971	0.0	2	20	45	KIAA1841 (k=51)
9	66	5.0	0.402	165	0	0	ENPP2 (k=20)
8	61	2.098	0.0	0	59	5	NDUFB1 (k=12)
14	61	2.066	0.0	0	22	41	TOM1L2 (k=20)

portant for transcriptional regulation, and is therefore an interesting gene to consider for further work.

Module 8 (dark green) is one of the two smallest modules with 61 nodes and a majority of the gene pairs have specific co-expression (Table 4.3). NDUFB1 is another new high-degree node in this module, and it is connected to all its 12 neighbors with specific co-expression. The gene encodes the protein called NADH:Ubiquinone Oxidoreductase Subunit B1, which is believed to transfer electrons in the respiratory chain from nicotinamide adenine dinucleotide (NADH) to ubiquinone [78]. *PRKAA1* is another gene of interest, as it has the highest centrality when it comes to closeness and betweenness. It encodes the catalytic subunit of an adenosine monophosphate (AMP)-activated serine/threonine protein kinase [78].

Module 4 (dark blue) and module 6 (blue) are closely connected, but divide the largest region of conserved gene pair connections in two. This indicates that this region might involve two different sorts of biological processes, and this was explored further with GO enrichment (in the following subsection). The network topology of module 4 and 6 are also slightly different. Even though they have almost the same number of nodes, module 4 has almost half the size of average degree and clustering coefficient compared to module 6 (Table 4.3). This means that module 4 is less connected to its neighbors and the neighbors are less connected with each other. Module 6 is the most highly connected of all modules, with a substantially higher average degree ($<k> \approx 6$) than the other modules, except module 9 which has a large average degree as well ($<k> = 5$).

Module 9 (turquoise) has the largest clustering coefficient ($C = 0.402$), which means that there is a higher tendency of the genes in this module to form tight clusters. Some of the clustered triangles can be observed directly in the turquoise module in the lower left of Fig. 4.5. All 165 gene pair connections in this module are of conserved type, and the node with the most of these connections is ENPP2. One of its neighbors, *FRMD4B*, is another potentially interesting gene, as it has the highest closeness- and betweenness centrality in the module. It encodes a FERM-domain containing protein with a likely role as a scaffolding protein [78].

Module 24 (dark yellow) has 90 nodes and most of these have S-type links between them. The highest connected node is the hub TMEM178A. Interestingly, it has specific co-expression with ANK1, which is the most central node according to betweenness and closeness in the module. It codes for the Ankyrin 1 protein, which attaches integral membrane proteins to the underlying cytoskeleton [78]. Previous studies have found this gene to be hypermethylated in AD [86, 87].

Module 3 (purple) is the most D-link dominated module. Here we find two of the network hubs, GTF2I and AQR. These have ZNF423 as a common D-linked neighbor, which is the modular node with highest betweenness centrality. *ZNF423* encodes a zinc finger DNA-binding transcription factor, which can both act as an activator or repressor [78]. Given its switch in correlation with two of the network hubs, this TF could be interesting to investigate in further studies. As indicated from the network visualization, module 3 overlaps with the two largest modules (indigo and green), but also with the smallest; the orange-colored module (module 14). Apart from one large bouquet with

the hub TOM1L2 in the core, the structure of module 14 is quite sparse. The hub LCAT is directly connected with TOM1L2 by differentiated (D) co-expression, but it has half the number of neighbors in the module.

Last, but not least, we have module 26, which is colored pink in Fig. 4.5. It has 68 nodes, of which as much as 52 represent the largest hub KIAA1841 and its first neighbors. All the other nodes in the module are therefore low-degree nodes. This is also the only module with all interaction types represented. In this module we find the only two gene pairs that have conserved co-expression among all the other S- and D-type of pairs in the differentially co-expressed region.

GO Enrichment Analysis of modules

Enrichment analyses of the genes within the modules were performed to look for significant functional associations to each module. Statistically significant GO categories (FDR < 0.05) were only found for three of the modules. All of these modules, namely module 4, 6 and 9, have exclusively conserved (C)-type of co-expression between their 92, 86 and 66 genes, respectively. We saw from Table 4.3 that these modules colored with different shades of blue were all homogeneous.

Table 4.4 shows some of the biological processes that are enriched for each of these three modules, sorted by fold enrichment (FE). For the sake of simplicity, only the most specific GO terms - of particular interest to AD - are shown. The complete results of all the significant terms are given in appendix B.2, Table 6.3-6.4. All processes are over-represented compared to what could be expected to be drawn randomly from the database of all human genes. Overall, we see many GO terms being *regulations* of other processes. This might indicate that the modules include some TFs, and are therefore important for transcriptional regulation.

Module 4 (dark blue) is mainly enriched in processes related to the development of the nervous system and synaptic activity (Table 4.4 and appendix Table 6.3). Several GO terms are related to signaling, transport and secretion by exocytosis, all of which are central processes in the synaptic vesicle cycle and for neurotransmitter secretion. A substantial amount of genes is enriched in neuron projection and regulation, which involves the development of extensions from the neurons, such as axons and dendrites. Also particularly interesting from a disease perspective is the regulation of long-term and short-term neuronal synaptic plasticity (FE = 45.83, $q = 1.20 \cdot 10^{-2}$, FE = 26.44, $q = 3.45 \cdot 10^{-2}$), which is essential to AD pathology.

Module 6 (blue) is most highly enriched in biological processes involved in amino acid metabolism (Table 4.4). The top 5 terms are related to the metabolism of the interchangeable amino acids glutamate, aspartate and oxaloacetate (Table 6.2 in appendix B.2). In fact, the module is fully enriched in glutamate metabolism by the two genes (*GOT1* and *GOT2*) responsible for producing either 2-oxoglutarate or aspartate from glutamate. GOT1 and GOT2 are cytosolic and mitochondrial aspartate aminotransferases, respectively, explaining the top 5 terms [78]. These are important regulators of glutamate levels, a major excitatory neurotransmitter in the CNS [4]. Although only two genes were enriched, GOT1 is one of the network hubs and therefore co-express with

Table 4.4: GO biological processes enriched in module 4, 6 and 9 ($N = 92,86,66$, respectively.) Only some of the most specific terms are included, sorted by fold enrichment (FE). #ref: number of genes in reference database. #genes: number of genes found in input gene list. FDR: Benjamini-Hochberg adjusted p-value.

GO biological process	#ref	#genes	FE	FDR
Module 4				
regulation of calcium ion-dependent exocytosis of neurotransmitter	3	2	>100	2.72E-02
regulation of synaptic activity	3	2	>100	2.70E-02
neurofilament cytoskeleton organization	8	3	85.92	3.43E-03
regulation of short-term neuronal synaptic plasticity	15	3	45.83	1.20E-02
regulation of vesicle fusion	23	3	29.89	2.67E-02
regulation of long-term neuronal synaptic plasticity	26	3	26.44	3.45E-02
associative learning	82	6	16.77	8.76E-04
vesicle docking	63	4	14.55	2.84E-02
negative regulation of neuron apoptotic process	152	6	9.04	1.24E-02
regulation of neuron projection development	522	18	7.90	3.60E-08
negative regulation of transport	449	9	4.59	2.52E-02
central nervous system development	1019	14	3.15	2.19E-02
Module 6				
glutamate catabolic process to 2-oxoglutarate	2	2	>100	1.61E-02
glutamate catabolic process to aspartate	2	2	>100	1.60E-02
mitochondrial ATP synthesis coupled proton transport	22	4	42.60	1.05E-03
phagosome acidification	28	4	33.47	2.09E-03
organelle transport along microtubule	80	6	17.57	5.44E-04
synaptic vesicle cycle	117	6	12.01	2.82E-03
respiratory electron transport chain	110	5	10.65	1.90E-02
regulation of macroautophagy	177	7	9.27	2.73E-03
cellular response to insulin stimulus	177	6	7.94	1.91E-02
regulation of exocytosis	211	6	6.66	4.02E-02
chemical synaptic transmission	414	9	5.09	1.30E-02
intracellular protein transport	992	13	3.07	3.89E-02
nervous system development	2203	21	2.23	4.50E-02
Module 9				
galactosylceramide biosynthetic process	6	3	>100	2.36E-03
central nervous system myelination	21	6	91.65	6.36E-07
oligodendrocyte differentiation	72	8	35.64	5.38E-07
peripheral nervous system development	77	5	20.83	4.07E-03
glial cell differentiation	180	9	16.04	1.56E-05
regulation of cell projection organization	710	10	4.52	2.97E-02
regulation of cellular component movement	1042	12	3.69	3.48E-02
regulation of hydrolase activity	1305	15	3.69	5.47E-03

several genes that might be functionally related. We also find processes related to transport and eradication of proteins, such as "organelle transport along microtubule" (FE = 17.57, $q = 5.44 \cdot 10^{-4}$) and "regulation of macroautophagy" (FE = 9.27, $q = 2.73 \cdot 10^{-3}$) (Table 4.4). These processes might be a response to the aberrant protein aggregation seen in AD patients. Finally, the terms involving most genes are nervous system development (FE = 2.23, $q = 4.50 \cdot 10^{-2}$) and intracellular protein transport (FE = 3.07, $q = 3.89 \cdot 10^{-2}$) (Table 4.4).

Module 9 (turquoise) is enriched in broad categories such as cellular development, both in the central- (CNS) and peripheral nervous system (PNS) (Table 4.4). More specifically, we find the generation of neurons, which includes both glial cell- and oligodendrocyte differentiation. Oligodendrocytes are large glial cells in the CNS, largely responsible for production of myelin (the lipid-rich insulating layer around neural axons) [88]. In fact, several of the enriched processes are related to myelination. The most specific and highly enriched GO term involves galactosylceramide, which is a sphingolipid composed of ceramide and a sugar unit. This is a key lipid in the composition of the myelin membrane [88]. In addition, as much as 15 genes are apparently involved in the regulation of hydrolase activity (FE = 3.69, $q = 5.47 \cdot 10^{-3}$), indicating that some of these genes might be TFs regulating the above-mentioned processes.

KEGG Pathway analysis of modules

Now with more local structures identified in the network, it was interesting to investigate which of these modules are responsible for the enriched pathways found in the network as a whole (Table 4.2). Several pathways in the KEGG 2019 database were found enriched in the modules 4 (dark blue) and 6 (blue), corresponding to the largest conserved region in the network. Table 4.5 and 4.6 show the significantly enriched pathways in module 4 and 6, respectively. Other than synaptic vesicle cycle (enriched in both module 4 and 6) and insulin signaling (enriched in module 4), the other four network pathways were not enriched in the modules. Nonetheless, new and more specific pathways showed up for modules that were not enriched when looking at the network as a whole.

Module 4 was enriched in two new and more specific pathways; GABAergic synapse (FE = 12.2, $q = 8.70 \cdot 10^{-3}$) and D-Glutamine and D-glutamate metabolism (FE = 87.0, $q = 0.0212$). The latter is a subpathway of the former, which explains why the two genes *GLS* and *GLS2*, coding for glutaminases, are enriched in both pathways. These enzymes convert glutamine to glutamate, which is then converted to gamma aminobutyric acid (GABA), the most abundant inhibitory neurotransmitter in the central nervous system (CNS) [4].

Module 6 had the largest number of pathways significantly enriched, and most of these can be related to AD. Only some of the most relevant terms are summarized in Table 4.6, while the full table is given in appendix B.4. Several of the pathways enriched in module 6, such as VEGF- and mTOR signaling are related to or involved in PI3K-Akt signaling, found earlier for the whole network. Module 6 was also highly enriched in pathways related to amino acid metabolism, the most highly enriched by far was the biosynthesis

Table 4.5: All significantly enriched KEGG Pathways in module 4, sorted by Fold Enrichment (FE). Overlap: ratio of genes found vs expected from the reference list. FDR: Benjamini-Hochberg adjusted p-value. Enriched genes are shown explicitly.

Term	Overlap	FE	FDR	Genes
D-Glutamine and D-glutamate metabolism	2/5	87.0	2.12E-02	GLS, GLS2
Synaptic vesicle cycle	6/78	16.7	4.87E-04	ATP6V0A1, DNM1, SNAP25, STX1B, STXBP1, SYT1
GABAergic synapse	5/89	12.2	8.70E-03	GABRA1, GLS, GLS2, GNG3, SLC12A5
Insulin signaling pathway	5/137	7.93	3.29E-02	BRAF, HK1, MAPK10, MAPK9, PRKAR1B

Table 4.6: Some of the significantly enriched KEGG Pathways in module 6, sorted by Fold Enrichment (FE). FDR: Benjamini-Hochberg adjusted p-value. Enriched genes are not shown for the sake of simplicity, but can be found here: 10.6084/m9.figshare.13344245.v2.

Term	Overlap	FE	FDR
Phenylalanine, tyrosine and tryptophan biosynthesis	2/5	93.0	5.07E-03
Synaptic vesicle cycle	7/78	20.9	2.70E-06
Epithelial cell signaling in Helicobacter pylori infection	6/68	20.5	2.40E-05
Oxidative phosphorylation	11/133	19.2	3.65E-09
Vibrio cholerae infection	4/50	18.6	2.16E-03
Parkinson disease	11/142	18.0	3.73E-09
Alzheimer disease	10/171	13.6	2.50E-07
Huntington disease	11/193	13.3	6.79E-08
VEGF signaling pathway	3/59	11.8	3.08E-02
Phagosome	5/152	7.65	1.19E-02
mTOR signaling pathway	5/152	7.65	1.29E-02
Cellular senescence	5/160	7.27	1.31E-02

of phenylalanine, tyrosine and tryptophan (FE = 93.0, $q = 5.07 \cdot 10^{-3}$). However, again only two genes (GOT1 and GOT2) were responsible for the over-representation (data not shown). We also find several pathways related to viral infection enriched, which indicates that this module might have a central role in immune responses. Interestingly, disease pathways are also enriched, such as Parkinson- and Huntington disease, which are two neurodegenerative diseases. Alzheimer disease was also highly enriched (FE = 13.6, $q = 2.50 \cdot 10^{-7}$), indicating that module 6 might be a disease module.

For the set of 66 genes in module 9 (turquoise), no significant enrichment was found in KEGG Pathways after multiple testing correction ($q = 0.077$). However, there was high fold enrichment in the two categories ether lipid- and sphingolipid metabolism (FE = 19.3, $p = 5.02 \cdot 10^{-4}$). Three genes were enriched in each category, of which *UGT8* and *GAL3ST1* were common to both. In addition, the intra-modular hub ENPP2 was enriched in ether lipid metabolism, and *CERS2* in sphingolipid metabolism. This pathway enrichment corresponds well with the enrichment in GO biological processes related to lipid metabolism (Table 6.4), shown earlier for this module.

Module 26 (pink) was the only module from the S/D-region which showed any significant enrichment after multiple testing. It was enriched with genes related to the Notch signaling pathway (FE = 24.5, $q = 6.54 \cdot 10^{-3}$). This pathway plays a key role in neuron development, and has been associated with neurological disorders [89]. The nodes enriched from this module were KAT2B, APH1B, HDAC1 and JAG1, the first three of which are nearest neighbors of and have differentiated (D) co-expression with the largest hub KIAA1841. This indicates that this novel transcript might have a regulatory role in signaling that is rewired as a response to AD. We will return to this topic in the discussion section.

Though not significant, Module 3 (purple) showed a 7-fold enrichment in cGMP-PKG signaling pathway ($p = 6.39 \cdot 10^{-4}$, $q = 0.197$), that might be interesting. The 5 genes *ATP1A2, ADRB2, PRKG1, SLC25A6* and the hub *GTF2I* were enriched. The remaining modules had no significantly enriched pathways ($q = 1.0$).

4.1.7 Biological functions of prominent genes

As only the most significantly co-expressed gene pairs end up in the CSD network, it would essentially be interesting to study each and every one of the genes and their interactions. However, since the network hubs represent genes that are co-expressed with a great number of other genes, exploring the biological function of these is especially interesting. The GeneCards database [78] was used as a starting point for functional annotation, followed by literature searches. A summary of the biological functions of all the 18 hubs can be found in Table 6.8 in appendix C, but the most important findings will be explored in the following paragraphs. We will start by describing the hubs of the S&D-rich region of the giant component, and then move on to the hubs of conserved regions. A visualization of the hubs and their first neighbors is given in Fig. 4.6, including which module each gene was assigned to in the modular decomposition. The numbering is in the same order as in the earlier shown table of hubs (Table 4.1).

KIAA1841 is by far the most highly connected node in the network, with 44 D-linked

and 25 S-linked interactions. It is centrally located in the largest region of the giant component, where it forms a bouquet-like (disassortative) topology with its neighbors. Apart from one hub, the majority of its neighbors are low-degree nodes. *KIAA1841* is a protein-coding gene which encodes a protein belonging to the KIAA-family of uncharacterized proteins, containing a domain of unknown function (DUF) [78]. To my knowledge, the biological function of this protein is unknown. It was therefore of interest to consider its neighbors to hypothesize its function based on the guilt by association principle. The gene set of the 69 neighbors of KIAA1841 showed no significant enrichment in any of the GO terms. KIAA1841 is however directly connected to the hub PLTP by an S-type link. *PLTP* encodes the Phospholipid Transfer Protein, which binds and transports a variety of lipid molecules, including cholesterol and vitamin E [78]. The specific co-expression (strong correlation only in AD samples) indicates a biological association that is disease-dependent, although the nature of which cannot be decided from this network alone.

MIGA2 is mostly connected by S-links, indicating condition-specific correlation with its neighbors ($k = 31$). It is quite disassortative, connecting only to one other hub (PLTP). The gene *MIGA2* encodes Mitoguardin 2, a protein located in the outer mitochondrial membrane, involved in glycerophosholipid metabolism [78]. Via phospholipase D6 (PLD6), it hydrolyzes cardiolipin into phoshatidic acid (PA) and phosphatidylglycerol (PG) [78].

TMEM178A is a fully homogeneous hub with only S-linked neighbors. *TMEM178A* encodes a transmembrane protein which is enriched in brain tissue [78]. It acts as a negative regulator of osteoclast (multinucleated bone cell) differentiation, which to my knowledge has unknown relation to neurogeneration. A transcriptional study did however reveal that TMEM178A was the largest dysregulated hub in the transition from normal to AD states [90], so it could be an interesting gene to investigate further. Other than the hub PLTP, it is directly connected to LCAT.

LCAT is the most heterogeneous of all hubs ($H = 0.5$), connecting to 11 nodes with specific (S) co-expression and 10 nodes with differentiated (D) co-expression. It codes for the enzyme Lecithin-Cholesterol Acyltransferase, which has a key role in cholesterol transport [78]. This protein is primarily located in plasma, but is also produced in the brain. The mix of S- and D-type links make the gene an interesting candidate for further studies.

AQR has a central position in the network and has almost exclusively differentiated co-expression with its neighbors. *AQR* codes for RNA helicase, which catalyzes the adenosine triphosphate (ATP)-dependent unwinding of RNA helices [78]. It is involved in pre-mRNA splicing as part of the spliceosome. Many cellular processes involve RNA processing, and thus a malfunctioning could potentially lead to disease. Defective alternative splicing has been associated with neurological disorders, including AD [91, 92].

AL158206.1 is a homogeneously D-linked hub ($k = 24$), and quite distanced from the other differentially co-expressed hubs of the network. AL158206.1 is a novel transcript coding for a long non-coding RNA (lncRNA), not a protein [78]. It has been proposed

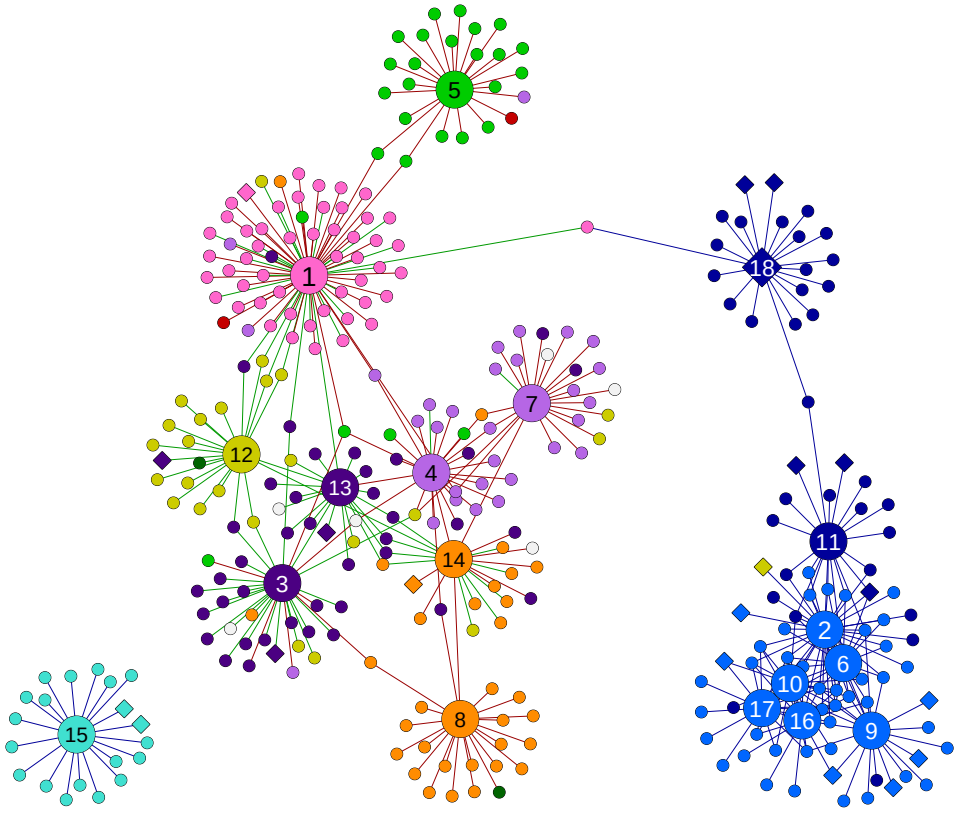


Figure 4.6: The 18 network hubs and their first neighbors. Hubs are enlarged nodes numbered from largest to smallest degree; 1: KIAA1841, 2: NMNAT2, 3: MIGA2, 4: AQR, 5: AL158206.1, 6: HPRT1, 7: GTF2I, 8: TOM1L2, 9: YWHAH, 10: GOT1, 11: NAPB, 12: TMEM178A, 13: PLTP, 14: LCAT, 15: ENPP2, 16: CADPS, 17: MDH1, 18: VSNL1. Colors of nodes indicate the module they belong to, using the same coloring scheme as earlier. Links are colored by co-expression type; blue = conserved (C), green = specific (S) or red = differentiated (D). Diamond nodes are previously AD-affiliated genes. $N = 339$, $M = 439$.

that lncRNAs have vital roles in transcriptional regulation, making this an interesting transcript for further studies [93]. The transcript also overlaps with an exon, which translates into the protein Alkaline ceramidase 2. This is a Golgi-localized enzyme involved in sphingolipid metabolism; it splits ceramide (a sphingolipid) into sphingosine and free fatty acids [78]. It is found to be up-regulated in response to DNA damage, where the increased sphingosine levels mediate programmed cell death [94].

TOM1L2 is also completely homogeneous, with D-links to all its 22 neighboring nodes, including the above-mentioned hub LCAT. *TOM1L2* translates into the Target Of Myb1-Like 2 Membrane Trafficking Protein. The protein belongs to a family of proteins involved in clathrin-mediated endocytosis, a form of vesicular transport [78]. Previous findings have associated clathrin-mediated endocytosis with APP trafficking, indicating its role in the production of Aβ in neurons [95].

GTF2I has 23 nearest neighbors, of which all but one of the interactions have strong correlations that switch sign when going from control to case (D-linked). *GTF2I* provides instructions for two different proteins; TFII-1 and BAP-135. The former is a general transcription factor, while the latter has been linked to the activation of B-cells (specialized white blood cells) in immune response [96]. TFII-1 is active in the brain and has been associated with the regulation of calcium flow into cells [97].

The hubs in the conserved regions are all homogeneous and most are highly interconnected, which explains the assortativity of the C-network (Fig. 4.4a). VSNL1 is a specially interesting hub because of its previous association with AD, although the hub has only C-type of interactions to its 20 neighbors. It codes for Visinin Like 1, which is a Ca^{2+}-sensor protein in nerve cells [78]. The protein regulates the activity of adenylyl cyclase, which modulates intracellular signaling in CNS. VNSL1 has a prominent position in the network, forming a connecting node between the S/D-rich- and the largest C-region (Fig. 4.6). Its neighbors SMYD2 and GLS, with specific co-expression with KIAA1841 and conserved co-expression with NAPB, respectively, are therefore also interesting candidates for further investigation ("bridge" node neighbors of node 18 in Fig. 4.6).

NMNAT2 is the second most connected node in the network and the largest of the conserved hubs ($k = 37$). It is a protein-coding gene which translates to Nicotinamide Mononucleotide Adenylyltransferase 2, a cytoplasmic enzyme predominantly expressed in the brain [98]. It transfers an adenylyl group from ATP to nicotinamide mononucleotide (NMN) to yield NAD+. This is a cofactor which is essential for a variety of cellular processes, which might explain why the gene is co-expressed with many neighbors with a conserved pattern.

Another hub related to nucleotide metabolism, and strongly connected with NMNAT2 in the network is *HPRT1*, which encodes Hypoxanthine Phosphoribosyltransferase 1 [78]. This protein is involved in the recycling of purines, which aside from forming DNA and RNA, are central components of important biomolecules such as ATP, cyclic-AMP (cAMP) and NAD [99]. Mutations in this gene is known to cause the neurodevelopmental Lesch Nyhan Syndrom, and the neurological aberrations resulting from the protein deficiency has been predicted to play a pathogenic role also in AD [99]. Some of the

consequences of HPRT deficiency are aberrant cell cycle control, DNA repair, membrane trafficking, defective neurotransmitters and sphingolipid metabolism [99].

The third largest conserved hub, YWHAH, has 22 nearest neighbors, including both of the forementioned largest C-hubs. *YWHAH* codes for the eta isoform of a protein which activates other proteins by binding to phosphoserine/threonine motifs. It is part of the 14-3-3 protein family that regulates many vital processes such as signal transduction, protein trafficking and apoptosis [78]. It has previously been associated with Schizophrenia, another neurological disorder [100].

Two hubs which are directly connected in the network and have closely related functional annotations are *GOT1* and *MDH1*. Both genes code for proteins involved in glucose metabolism, more specifically the tricarboxylic acid (TCA) cycle [4]. *GOT1* codes for the cytosolic form of Glutamic-Oxaloacetic Transaminase 1 [78]. GOT2, the mitochondrial form, also exists in the network, but its only neighbor is MDH1. *MDH1* encodes Malate Dehydrogenase 1, a cytosolic enzyme which catalyzes the reversible oxidation of malate to oxaloacetate [78]. It is NAD-dependent, which might explain why the gene is also co-expressed with NMNAT2 (which as mentioned provides NAD+).

Directly associated in a cluster with GOT and MDH1 is *CADPS* (Fig. 4.6). This gene is strongly enriched in brain tissue, where it codes for the Calcium-Dependent Secretion Activator, a membrane protein associated to synaptic vesicles [101]. It is potentially a Ca^{2+}-sensor which triggers the release of neurotransmitter from vesicles by exocytosis [101]. Functionally related to CADPS, but not directly connected in the network, we find NAPB (module 4 in Fig. 4.6). *NAPB* encodes the beta subunit of the soluble N-ethylmaleimide-sensitive factor (NSF) Attachment Protein Receptor (SNARE) complex [78]. The protein complex is involved in vesicle-mediated transport between the endoplasmic reticulum (ER) and the Golgi apparatus. It is also preferentially expressed in brain tissue [78].

Lastly, ENPP2 is the hub of the disconnected C-linked graph (turquoise module), and therefore quite distanced from the other hubs (Fig. 4.6). The gene, also referred to as *NPP2*, encodes autotaxin, a member of the nucleotide pyrophosphatase and phosphodiesterase family [78]. Additionally, it has a lysophospholipase D activity, and this generation of lysophosphatidic acid (LPA) stimulates cell proliferation [102]. LPAs are phospholipids that have been implicated in AD, but their role in the potential pathology is unknown [103].

Genes previously associated with AD

A systematic search for genes in the CSD network affiliated with AD was done using the MalaCards integrated database [25]. From the original population of 21044 genes in the microarray data, 420 genes had previous association with AD. As much as 64 of these genes were recognized in the network of sample size $N = 1535$. This is more than expected to be drawn randomly (FE = 2.09, $p = 1.49 \cdot 10^{-8}$). All the AD-related genes identified in the CSD network are listed with the modules they belong to in Table 4.7 and visualized in the network in Fig. 4.7. In the figure they are highlighted by enlarged, diamond-shaped nodes. Following the guilt by association-principle, essentially all of

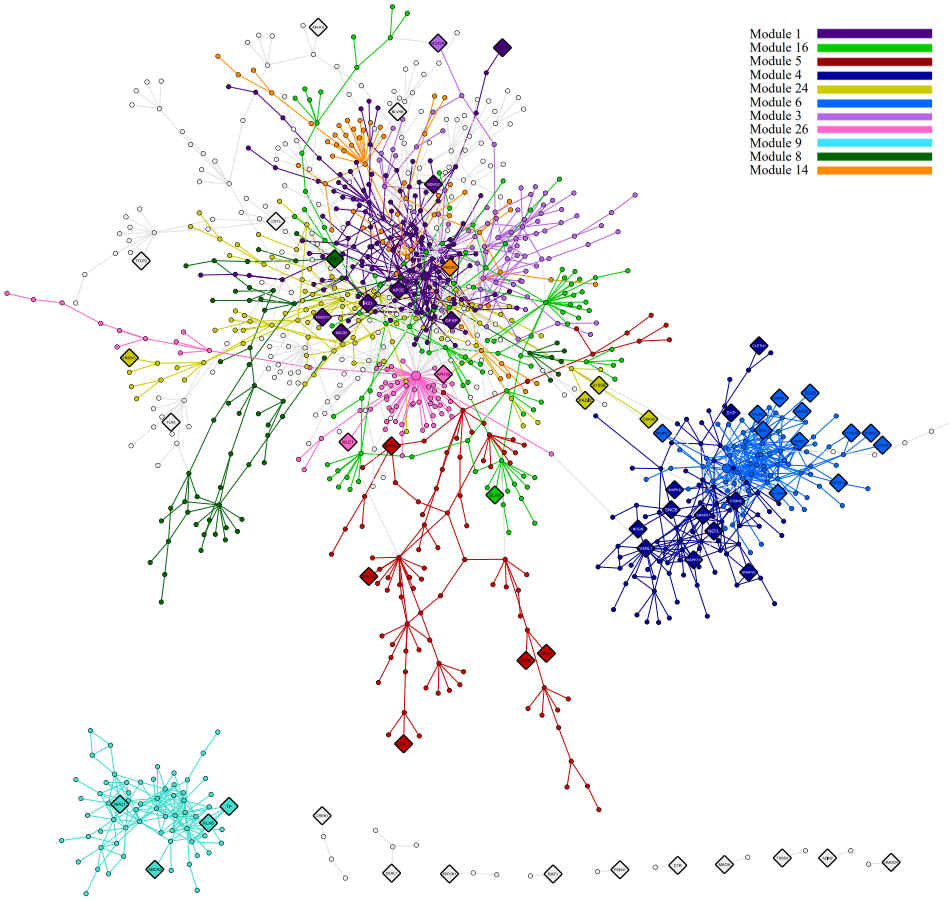


Figure 4.7: The 64 genes previously associated with AD (diamond nodes) recognized in the CSD network. Genes (nodes) are colored according to the module they belong to, other than the genes outside of the modules, which are gray. $N = 1169$, $M = 1816$.

the nearest neighbors of these nodes would be interesting to consider further. However, due to time constraints only some will be considered here.

Overall from Fig. 4.7 we see that the AD-related genes are distributed across the whole network, but with a larger concentration in the conserved (blue) regions. In fact, module 4 and module 6 have the largest number of disease-associated genes, 11 and 12 genes respectively (Table 4.7). This is more than 1/3 of the gene list, an over-representation compared to what can be expected by chance (FE = 6.47, $p = 1.30 \cdot 10^{-12}$). These modules are therefore candidate disease modules, although the gene pair correlations are conserved between the two conditions.

Another observation from Fig. 4.7 is that many of the AD-related genes are peripheral nodes in the network with low degrees. More than half of the 64 diamonds have degree

$k = 1$, meaning that the genes are only strongly co-expressed with one other gene. 16 genes are not part of the large modules identified in the module analysis ("Other" in Table 4.7). This might be explained by the degree distribution of the network, where most genes are in fact low-degree nodes. Interestingly though, all conserved hubs have at least one AD-related neighbor (see the blue nodes and their diamond shaped neighbors in Fig 4.6). For example, YWHAH is co-expressed with as much as four AD-related genes (*CALM3, SNCA, UCHL1* and *VDAC1*), making it a prominent candidate for further studies. *VSNL1* is the AD-related gene with highest degree, and the only one of the hubs. It is the intra-modular hub of module 4, which further indicates that this module is a potential disease module. Note-worthily, VSNL1 is connected to SMYD2 (C-link), which again is connected to KIAA1841 (S-link) and has one of the top 10 highest betweenness centralities in the giant component. *SMYD2* encodes the N-lysine methyltransferase, involved in cell proliferation and cancer [104]. Its annotated function and prominent position in the network makes *SMYD2* a novel interesting candidate for further research.

Several genes in the network, some of which were already associated with AD, encode parts of the large ATP Synthase complex. ATP5F1A and ATP5PD, both located in module 6, are previously AD-associated (Table 4.7). Directly co-expressed with ATP5F1A we find ATP5F1B and ATP5MC3, all three forming a cluster of conserved co-expression in the module. All genes (starting with ATP5) encode different subunits of the mitochondrial ATP synthase [78]. ATP5F1B is also co-expressed with the AD-related gene *SLC25A4*, which encodes a solute carrier protein exchanging cytoplasmic ADP with mitochondrial ATP across the inner mitochondrial membrane [78]. Based on the guilt by association principle, ATP5F1B and ATP5MC3 are potential disease genes. Also part of the clustered region are the hubs GOT1 and MDH1, which as mentioned earlier have related functions. Together, the mentioned genes are largely responsible for the enrichment in GO terms related to mitochondrial ATP synthesis (Table 6.2) and the KEGG pathway Oxidative phoshorylation (Table 4.6) in this module.

In relation to oxidative phosphorylation, several genes in module 6 also encode vacuolar ATPases (v-ATPases), with ATP6V1A being especially interesting due to its association with many neighbors ($k = 17$). Among its neighbors we find many of the hubs (NMNAT2, HPRT1, YWHAH, CADPS, GOT1 and MDH1), but also the two AD-related genes *UCHL1* and *STMN2*. In neurons, V-ATPases have been shown to generate a proton-gradient in synaptic vesicles that provides the energy for loading and release of neurotransmitters [105]. This supports why the *ATP6*-genes were enriched in the synaptic vesicle cycle in addition to oxidative phosphorylation (data not shown).

Also interesting are the three AD-related genes of module 24 (dark yellow diamonds in Fig. 4.7), located on the border between the two separated regions of the giant component: GSK3B, S100B and PADI2. GSK3β (Glycogen Synthase Kinase 3 Beta) is a serine/threonine kinase involved in many essential pathways and with a central role in AD, mainly through its involvement in the phosphorylation of protein Tau [106]. S100B is a calcium binding protein which is associated with AD through its ability to suppress Aβ aggregation [107]. *PADI2* encodes Peptidyl Arginine Deiminase 2, a Ca^{2+}-dependent enzyme that converts arginine to citrulline on substrate proteins, including myelin ba-

Table 4.7: Genes previously associated with AD and their location in the modules of the CSD network, sorted from largest to smallest module. 16 genes (Other) were found outside of the 11 modules analyzed.

Module	AD-affiliated genes
1	APOE, CSF1R, GFAP, IRS1, MAOB, NDRG2, TMED10
16	GLRX
5	DVL1, EIF2S1, MTR, PLA2G6, SPHK2
4	CLSTN1, MAPK10, MAPK9, NEFL, NRGN, SNAP25, SNAP91, SNCB, STMN2, SYP, VSNL1
24	GSK3B, HMOX1, PADI2, S100B
6	ATP5F1A, ATP5PD, CALM3, COX5A, COX6C, MAP2K1, MARK1, PLD3, SLC25A4, SNCA, UCHL1, VDAC1
3	SLC25A6
26	APH1B, KLC1
9	ABCA2, KLK6, NPC1, TF
8	CAT
14	ABCA7
Other	APH1A, AZIN2, BLVRB, C1R, CREB1, CST3, DERL1, GLRX, HMGCR, MAOA, PPP3R1, PSEN1, RAF1, TFCP2, THBS4, TLR2

sic proteins in the CNS [78]. Of the three, GSK3B has the most central location in the network, forming a bridge over to NMNAT2 in module 6. S100B and PADI2 have only one neighbor, but it is a node shared by all three disease genes: GRSF1. GRSF1 is connected to S100B and PADI2 by a differentiated (D)-link and to GSK3B with a specific (S)-link. It also has specific (S) co-expression with an AD-related gene of module 6 (*MARK1*). GRSF1 has a relatively high betweenness centrality, which suggests that it is important for the information flow in the network. Since it is co-expressed with four AD-related genes, guilt by association highly implies *GRSF1* to have a role in the disease. GRSF1 (G-rich Sequence Binding Factor 1) is an RNA-binding protein (RBP) required for posttranscriptional mitochondrial gene expression [78, 108].

Ultimately, it is worth mentioning that 7 AD-related genes were found in the largest module (module 1, indigo) in the network (Table 4.7). Of these, we find APOE (Apolipoprotein E), the infamous gene of which the ϵ4-allele is associated with increased risk of AD [25]. It has a modest role in the network, connecting only to MIGA2 by an S-link. This however further supports that *MIGA2* might have a role in AD. Also interesting is *IRS1*, which has specific (S) co-expression with nine other nodes, among them the hub PLTP. This gene encodes IRS1, a protein substrate of the intracellular insulin receptor (IRS), which when tyrosine-phosphorylated activates PI3K [78]. Decreased phosphorylation or amounts of proteins in the insulin-IRS1-Akt pathway has been observed in AD brains, and this defective insulin signaling leads to synaptic dysfunction and impaired memory [109, 110].

4.2 Integrative Analysis

This part complements the already established CSD framework in the hope of discovering new interesting features of the network. To enhance the interpretation of the complex differential co-expression network and better predict biological functions, two main strategies were employed. First, differential mean expression levels were used to increase the knowledge of regulation on the level of individual genes. This was initially performed on all 21044 genes in the transcriptomic data, and then integrated on top of the CSD network. Second, new data in the form of PPIs were added as a network layer to look for signals on the protein level. There is often a need to integrate networks at different molecular levels (e.g. transcriptome, proteome) to fully understand the link between gene regulation and a resulting phenotype (in this case AD) [55].

4.2.1 Differential Expression Analysis

Differential expression analysis was performed to reveal potential disease-associated genes not found in the differential co-expression network, and more importantly to strengthen the signal of those actually forming part of the network. The resulting list of genes that have changed mean expression between healthy and AD-individuals can contribute to further insight into essential genes and processes related to the disease.

The overall result of the differential expression analysis on the AD microarray data (80 diseased and 93 age-matched normal controls) is visualized in a volcano plot (Fig. 4.8). In this plot, the negative of the $\log_{10}$-transformed FDR-values (significance) is for each gene in the data set plotted against the $\log_2$FC (magnitude). The vertical and horizontal lines represent the "double"-filtration (Fig. 4.8). Significantly up-regulated genes (up-DEGs) are colored red and down-regulated genes (down-DEGs) blue. In total, 1196 genes were differentially expressed ($q = 0.05$), of which 699 were down-DEGs and 497 were up-DEGs. The complete list of up-regulated and down-regulated genes can be found in the following doi: 10.6084/m9.figshare.13366061.

In the plot (Fig. 4.8), the top 10 genes with the highest expression changes are labeled. The summary of these top five up- and down-regulated DEGs with their expression changes and statistical parameters is shown in Table 4.8. The top up-DEGs - *RGS1*, *CD163*, *LINC01094*, *ADAMTS2* and *HLA-DRA*, and the top down-DEGs - *SST*, *BDNF*, *PCDH8*, *MIR7-3HG* and *CALB1*, are all good candidates for further studies. The elevated or decreased abundance in mRNA levels of these genes might lead to similar changes in protein levels, finally affecting their respective biological processes and potentially influencing disease transition. Further investigation is needed to validate the results and explain the biological roles of the genes. Moreover, we see that the overall magnitudes of change in mean expression between healthy and sick individuals were quite low (Fig. 4.8). In fact, no genes have $|\log_2 FC| > 1$, which is a commonly chosen threshold in the literature for analyses of differential expression. To investigate if the low values might be due to tissue specificity, DEA was also performed on each of the four brain regions (see appendix G). The analysis suggested that the hippocampal region is more affected by AD, but due to time constraints this was not explored further.

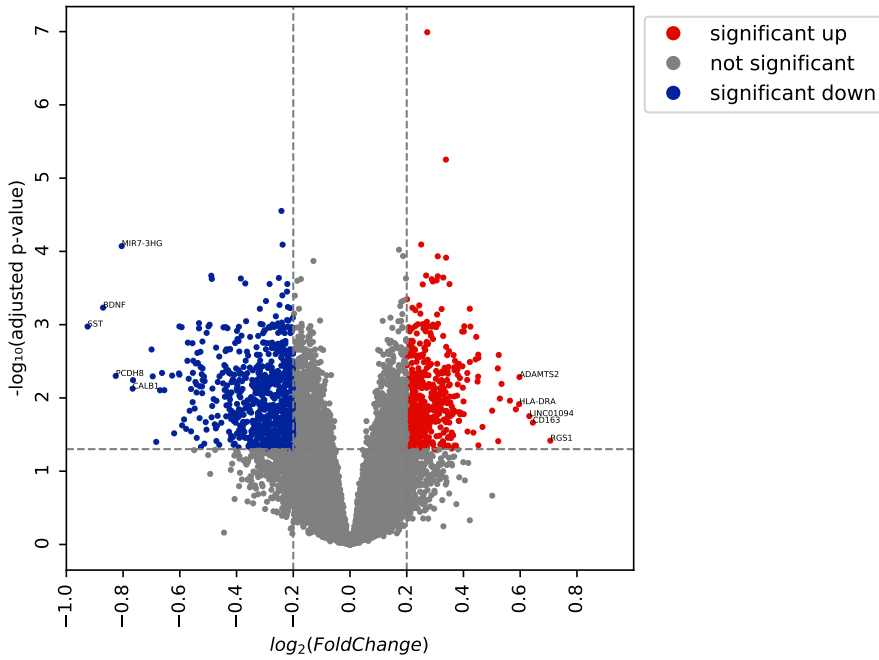


Figure 4.8: Volcano plot of average gene expression changes in AD vs control in terms of $\log_2$ fold-change (x-axis) and $-\log_{10}$FDR-corrected p-value (y-axis). The most up-regulated genes are towards the right (red), the most down-regulated genes are towards the left (blue), and the most statistically significant genes are towards the top. Genes with $|\log_2 FC| > 0.2$ and $FDR < 0.05$ are considered significantly differentially expressed (DEGs). The top 10 genes with greatest absolute change are labeled with gene symbols.

Table 4.8: Top 5 up-DEGs and down-DEGs among all brain tissue of individuals >= 60 years in Alzheimer's dataset, sorted by $\log_2$FC ($\log_2$ Fold Change). Mean gene expression is transformed with logarithm base 2 ($\log_2$). AD: Alzheimer's disease. FDR = Benjamini-Hochberg adjusted p-value.

Gene name	Control exp.	AD exp.	$\log_2$FC	Raw p-value	T-statistic	FDR
RGS1	6.47	7.17	0.71	4.31E-03	2.89	3.83E-02
CD163	7.60	8.24	0.64	1.68E-03	3.19	2.18E-02
LINC01094	7.08	7.72	0.63	1.18E-03	3.30	1.77E-02
ADAMTS2	6.35	6.94	0.60	1.34E-03	3.91	5.19E-03
HLA-DRA	9.26	9.85	0.60	6.07E-03	3.49	1.22E-02
CALB1	7.45	6.68	-0.77	2.66E-04	-3.72	7.47E-03
MIR7-3HG	6.89	6.09	-0.80	2.42E-08	-5.85	8.47E-05
PCDH8	8.67	7.84	-0.83	1.23E-04	-3.93	5.01E-03
BDNF	6.42	5.55	-0.87	1.11E-06	-5.05	5.86E-04
SST	8.41	7.49	-0.93	3.98E-06	-4.77	1.06E-03

In order to infer some biological meaning of the DEA, GO enrichment analyses were performed separately for the 497 up-DEGs and 699 down-DEGs. Separate enrichment analyses of biological processes and pathways for up- and down-regulated genes has been suggested by Hong et al. to be more powerful than analyzing all DEGs together [111]. The lists of the most highly enriched processes are added to appendix B.3. Several of the up-regulated processes are related to immune responses and related signaling pathways. The most enriched process includes all genes in the reference, which was "peptide antigen assembly with MHC class II protein complex" (FE = 42.90, $q = 1.79 \cdot 10^{-3}$). The down-DEGs were mostly enriched in processes involved in the regulation of the synaptic vesicle cycle, including vesicle priming, -docking, -recycling and -transmission. The largest fold enrichment found for down-DEGs was in "maintenance of presynaptic active zone structure" (FE = 29.45, $q = 6.44 \cdot 10^{-3}$).

DEGs in CSD network

To inspect the location of DEGs in the CSD network, the $\log_2$FC node attribute was studied in Cytoscape. In total, 350 genes in the network had gene expression changes with an adjusted p-value lower than the threshold of 5% false positives. This means that, statistically speaking, less than 18 genes are falsely identified as differentially expressed. However, some of these genes had very low magnitudes of change, and therefore cannot be justifiably called DEGs. The value of $\log_2 FC = \pm 0.2$ was chosen to define a node in the network as a DEG ($q = 0.05$, $p = 6.6 \cdot 10^{-3}$), indicating at least a 15% change in expression from control to case on average over all samples. The location of these DEGs are highlighted in Fig. 4.9 and the gene names given in Table 4.9, where they are sorted by the main regions visible from the figure. The nodes in the CSD network are only colored red/blue if the magnitude of change was above the threshold and the differential expression significant (FDR < 0.05), following the same coloring scheme as in the volcano plot (Fig. 4.8). Some nodes are omitted from the visualization in order to highlight the DEGs, and hence the network has a lower size than the original in Fig. 4.1. Genes with previous association to AD are diamond-shaped (Fig 4.9) and marked in bold (Table 4.9). Both graphics can be combined for keeping up with the following paragraphs.

An important question is if the DEGs are enriched in the network, more than expected from the original microarray expression set. Significant enrichment would suggest that the network has an association with AD progression at the level of transcriptional regulation. In fact, 229 nodes from the original 1535 nodes in the CSD network were recognized as DEGs. One would expect only about 87 DEGs by chance, so this was a 2.6-fold enrichment ($p = 5.88 \cdot 10^{-44}$). Hence the CSD network is enriched with genes showing individual differential expression, in addition to their gene pair correlated expressions. All of the 229 DEGs would be interesting candidates for further research, thus the complete list is given in appendix E.

The subset of genes identified from DEA showed significant, although quite weak signal differences (low magnitude of change) between the two conditions. The gene with the greatest change in expression recognized in the network was *ADAMTS2* ($\log_2$FC ≈ 0.60, $q = 5.19 \cdot 10^{-3}$). This was the only one out of the top 10 DEGs in the microarray data (Table 4.8). The gene encodes a member of the ADAMTS family of disintegrin and met-

allopeptidases with thrombospondin motifs [78]. The gene is up-regulated in AD and has a central position in the S&D-region of the CSD network (the darkest red node in the middle of Fig. 4.9). ADAMTS2 is located in the largest module (indigo) and has ten neighbors ($k = 10$). It has an S-link to another up-regulated gene, namely ANGPT1 ($\log_2$FC ≈ 0.43, $q \approx 0.03$). Another neighbor is the AD-related gene *GFAP* (red diamond in Fig. 4.9), to which it has a specific (S) co-expression pattern. Although ADAMTS2 is not one of the previously AD-associated genes, another gene in the same protein family (*ADAMTS4*) has been associated with the disease. ADAMTS4 cleaves brevican, a CNS-specific protein suggested to be important in neuroprotection [112].

A striking feature is that all of the DEGs in the largest conserved region on the right side in Fig. 4.9 (corresponding to module 4 & 6) are down-regulated and closely connected. Module 4 and module 6 have 31 and 55 down-DEGs, respectively, together representing a more than 3-fold enrichment of DEGs in the network ($p = 1.17 \cdot 10^{-30}$). The expression of these genes are on average slightly lower in the individuals with AD compared to controls. Six of the down-DEGs in this region are the homogeneously conserved hubs NMNAT2, CADPS, YWHAH, HPRT1, GOT1 and NAPB. These are significantly down-regulated in AD, but to a varying degree ($\log_2$FC from -0.48 to -0.32, $q = 0.039$).

Oppositely, the DEGs in the S&D-region of the network are more dispersed and of both signs of fold change. These genes are both differentially expressed individually and differentially *co*-expressed (colored nodes with green or red links between them in Fig. 4.9). TMEM178A was the only differentially co-expressed hub recognized as a DEG, being slightly down-regulated ($\log_2$FC = -0.204, $q = 0.026$).

Interestingly, in the next largest conserved region (module 9) all DEGs are up-regulated, although to a lower extent. These are located on the left side of the module (Fig. 4.9). The other up-DEGs in the CSD network seem to be quite spread out in the network, and not clearly localized to any particular modules.

In addition to *GFAP*, some other DEGs in the network were also previously associated with AD. These are represented as diamond nodes in Fig. 4.9 and marked in bold in Table 4.9. The three most differentially expressed AD-related genes in the network were STMN2 ($\log_2$FC = -0.473, $q = 0.026$), SNCB ($\log_2$FC = -0.413, $q = 0.034$), and SYP ($\log_2$FC = -0.402, $q = 0.004$). These three are all located in module 4. Although most of the DEGs in the network were not previously associated with AD, it cannot be ruled out that similar genes are found in the MalaCards database. One example of this is the already mentioned *ADAMTS4*.

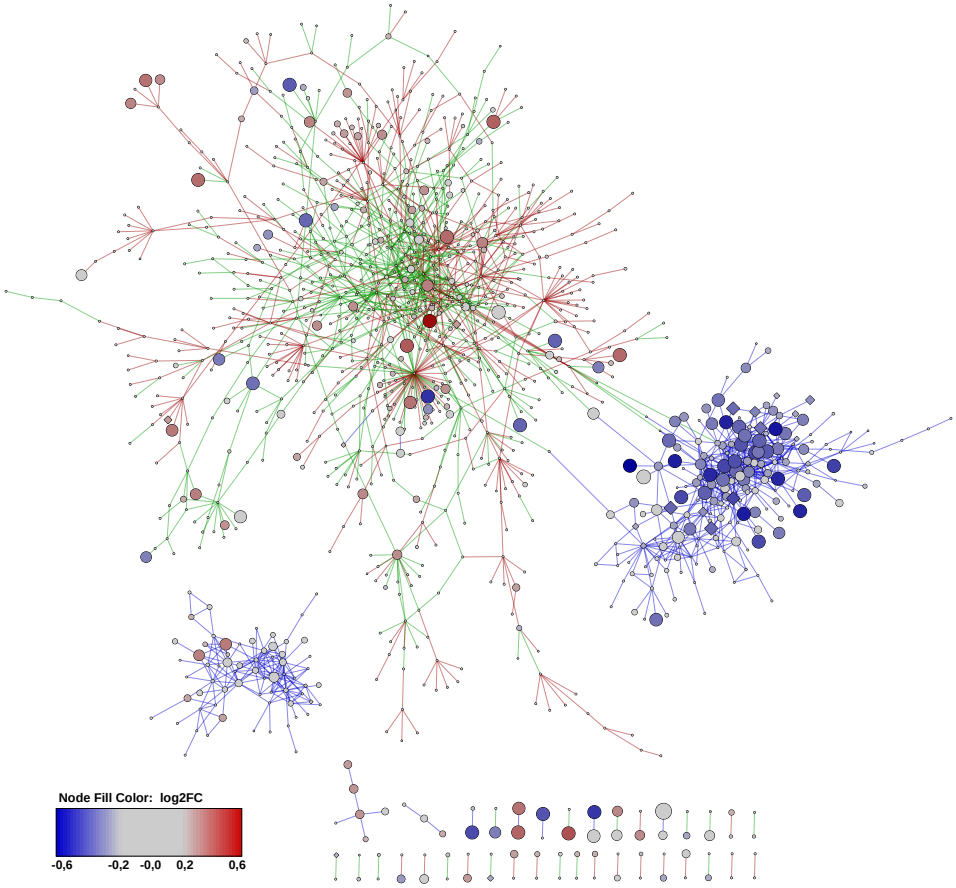


Figure 4.9: DEGs in the CSD network. Node size corresponds to the magnitude of change in mean gene expression ($|\log_2 FC|$). Colored nodes (DEGs) are above the threshold $|\log_2 FC| > 0.2$ AND significant after multiple testing correction (FDR < 0.05). The fill color is mapped by the sign of $\log_2 FC$ (see bottom-left chart); red and blue are up-regulated (+) and down-regulated (-) genes, respectively. The larger $|\log_2 FC|$ the darker the color and the larger node size. Links are colored by co-expression type; blue = conserved (C), green = specific (S) or red = differentiated (D). Diamond nodes are previously AD-associated genes. $N = 1219$, $M = 1841$.

Table 4.9: All DEGs recognized in the CSD network. Genes are categorized by region and whether they are up-regulated (Up) or down-regulated (Down). All genes are listed from the largest to the smallest magnitude of change (absolute $\log_2$FC). Genes previously associated with AD are marked in bold. C-region: DEGs within region of conserved co-expression, belonging to module 9 (only up-DEGs), 4 or 6 (the latter two only down-DEGs). S&D-region: DEGs within the specific and differentiated-linked region of the giant component. Other: The rest of DEGs outside the giant component, bottom of Fig 4.9.

Network Region	Up/Down	Gene symbol
C-region	**Up**	DOCK5, HSPA2, RNF130, RDX, USP54, NDE1, FRYL, GPRC5B, VAMP3
	Down	RGS4 MAL2 OLFM3 KCNV1 RAB3C CDC42 NAP1L5 GABRG2 SYN2 NMNAT2 **STMN2** CADPS SYT13 C3orf80 GNG3 ATP6V1G2 SEZ6L2 BEX5 PAK1 YWHAH NSF **SNCB** HPRT1 SYNGR3 **SYP** AMPH KALRN ACOT7 **UCHL1** SYN1 DYNC1I1 ATP6V1B2 ATP8A2 EEF1A2 NECAP1 AP3B2 SCN2A TMEM178B GOT1 **CALM3** SV2B SYT1 MLLT11 BEX1 DNM1 **PLD3** PGAP4 SYNGR1 STX1B PHF24 NAPB TAGLN3 GPRASP1 MOAP1 ENO2 GPRASP2 SCG5 ADAM23 TBC1D9 ATP6V1A ITFG1 STXBP1 NDRG4 MAPRE3 EID2 DNAJC5 RNF41 FBLL1 INPP5F SV2A GPI SCAMP5 SULT4A1 **VDAC1** CD200 AP2M1 MRPL15 GLT1D1 ATP5MC3 ATP6V1E1 BTBD10 ATP5F1B TOMM20 RAN PEX11B APOO
S&D-region	**Up**	ADAMTS2 ANGPT1 VAC14-AS1 HLA-DRB1 ITPKB GMPR NFKBIA CD74 SELL ITGB8 HCLS1 BBOX1 S1PR3 DNALI1 NACC2 PRKX AC005332.4 H1-2 RFX4 H2BC19P CSRP1 TP53INP1 CHST6 C1orf87 ZFP36L2 CGNL1 ZCCHC24 MAP4K4 **GFAP TLR2** CHD7 MTMR10 RIN2 PANTR1 KIAA1958 EZR LIFR CCDC69 PLEKHA7 EMX2OS DIAPH3 CTSH ANO6 AFF1 CRB1 ADGRA3 PTPN21 ZNF423 ARHGAP42 CRB2 HIPK2 MYO10 TEX26 STEAP3 GOLIM4 CDC42EP4 **KLC1** PLXNB1 TOB1 SERPINI2 HMG20B TMEM47 CCDC151 IKBKB HEATR5A IQCK RNF19A PRKG1 NXT2 OR7E14P RFX2 CXCL16 CERS1 PLXNB2 PELI2 ANAPC16 TAB2 NOTCH2 TNS2 HDAC1 FBXL7 PTBP1
	Down	RTN4RL2 PCLO TCERG1L HSPB3 CPNE4 RNF128 JPT1 UBE2T CLSTN2 CLSTN3 RFPL1S KLF10 LYNX1 SLITRK3 AC139256.2 AFF2 BCAT1 AL031118.1 KIF3A AC005229.4 AC006058.1 PEX3 RER1 ADAM11 TMEM178A WDR47 CCDC32 WDR74
Other	**Up**	MYBPC1 HLA-DRB1 CD74 SLC38A2 C5AR1 ZCCHC24 MAP4K4 CHD7 POU3F2 ST6GALNAC3 ANP32B MMP8 C21orf62 HIPK2 LRP4 GLIS3 RNF19A USP53 WDR49 SAP30L
	Down	GAD1 BRWD1 SNX10 LYRM9 RASAL1 RUNDC1 SGIP1 AC139256.2 **HMGCR** TFRC NKD2 YWHAZ PEX3

4.2.2 Protein-Protein Interactions

A comparison with the reference PPI network was done to determine whether the edges of the CSD network could represent a physical interaction between the proteins resulting from the expression of the pair of genes. The HI-Union data set downloaded from CCSB [84] contained 9094 genes with 64 006 interactions (of which 764 were self-loops). Of these genes, 8103 (89 %) were present in the original microarray data. However, as this transcriptomic study included 21044 genes, only around 39 % of these were represented with at least one protein interaction in the PPI network. Of the 1535 genes present in the CSD network, 713 (46%) had at least one connection to another protein in the PPI network. This is a slight over-representation (by factor 1.21, $p = 2.79 \cdot 10^{-11}$).

Three gene pair interactions were found in common between the PPI and CSD networks, with the type of co-expression indicated in between the gene names:

- NAPB (C) SYT1
- ATP5F1A (C) ATP5F1B
- SDCBP (D) TMEM17

The first PPI found was between NAPB and SYT1, located in module 4 (dark blue). The C-link between the nodes indicates a strong correlation in their gene expressions in both conditions. NAPB is already mentioned as a network hub, with its 21 C-linked neighbors. SYT1, however, has a more modest role in the network ($k = 3$). The gene codes for Synaptotagmin-1, a protein in the membrane of synaptic vesicles thought to work as a calcium sensor that triggers neurotransmitter release by exocytosis [78]. *SYT1* was also one of the genes enriched in the synaptic vesicle cycle pathway in KEGG (see Table 4.5). Their conserved co-expression and protein-protein association indicate that they are both part of this pathway, whether they physically bind or not. The SNARE complex that *NAPB* encodes a subunit of is known to contribute to membrane fusion necessary for exocytosis, which again is dependent on calcium [113].

The second interaction was between ATP5F1A and ATP5F1B, encoding two subunits of the same protein complex, which explains why they were connected in the PPI network. The genes have conserved co-expression, indicating that the two subunits of the mitochondrial ATP synthase complex are regulated similarly in both conditions. ATP5F1A was earlier identified as an AD-related gene, however. Both are down-regulated in AD according to the DEA, although only ATP5F1B significantly (log_2FC = -0.216, $q = 0.026$). The identified PPI supports the assumption of the two down-regulated genes forming part of a strongly connected complex. This again indicates a certain downregulation of ATP synthesis and oxidative phosphorylation in AD.

The last overlapping pair was SDCBP and TMEM17, which has differentiated (D) coexpression in the CSD network. The gene pair is found in the periphery of the network in module 5 (red). TMEM17 is a one-degree node, while SDCBP is connected to several other one-degree nodes forming a bouquet-like structure. *SDCBP* encodes Syntenin-1, a protein involved in vesicular trafficking that has several PDZ domains that bind a variety of transmembrane proteins [78]. *TMEM17* encodes such a transmembrane

component, which is localized to the cilia transition zone, where it is required for sonic hedgehog (SHH) signaling, a pathway involved in nervous system development [78, 4]. The differentiated type of co-expression suggests that the interaction is present under both conditions, but with possibly different underlying mechanisms.

Chapter 5

Discussion

This thesis aimed to apply network analysis in the search for genes and biological processes involved in AD pathology. The CSD framework was the foundation for the comparative study of gene expression from patients with AD compared to healthy individuals. This method successfully constructed a differential co-expression network, showing three types of transcriptional correlation between gene pairs. Although great insight can be drawn from the network, it was an overall challenge to infer clear biological patterns. This might be due to the disease's genetic complexity and a multitude of environmental factors affecting its development. The following sections will highlight and discuss some of the main results of this comparative study, closing with an elaboration of challenges and method limitations. We will start with evaluating the overall network properties for verification of the CSD method and the resulting biological predictions.

5.1 Overall network analysis

5.1.1 Topological properties

The topological overview of the CSD network indicates a non-randomness in the organization of genes. A power-law degree distribution was observed, which was expected from earlier studies on co-expression networks [114, 115]. This suggests that the topology of the network is not random, but rather that the gene connections result from real biological relationships. However, it is important to recall that the links in this network do not reflect any direct biological interactions [3]. In addition, the CSD method was applied to gene expression samples extracted from individuals at one particular time point. Therefore, the resulting network is static, while in reality, gene expression is a dynamic process. For these reasons, we need to be careful about predicting or concluding too much from the network.

The study of the overall topology of the different co-expression networks showed some

interesting results. The individual C-network had quite different topological features compared to the S- and D-type networks, which might be due to different underlying regulatory mechanisms. The genes with conserved type of co-expression were densely connected, and the overall C-network was slightly assortative. This demonstrates a tendency of positive degree-degree correlation, which was supported by most of the nine "conserved" hubs directly connecting to one or more of the other hubs of the same type. On the other hand, the specific and differentiated networks show disassortative topologies with negative degree correlations. The hubs within the S/D-region were generally not linked to each other directly, but through intermediate low-degree nodes. These degree correlation patterns are quite similar to the ones found in the original article describing the CSD method [3]. Even though there was an overall tendency in the CSD network of same type of co-expression to group together, this homogeneity was more evident for the conserved type of genes. These were highly homogeneous, with only 9 genes also being connected to either S- or D-type links. The S-type and D-type of interactions were more overlapping in the giant component of the CSD network, and some of the hubs were quite heterogeneous. It is difficult to determine a definite cause behind these observations.

Further analysis of modular structures and central genes was carried out to give more insight into the underlying biology of the CSD network. The Louvain community detection algorithm successfully partitioned the network into modules, verified by the high modularity score ($Q = 0.83$). Only the 11 largest modules (size above 50) were chosen for further studies. This was considered both manageable and sufficient for downstream analyses. Interestingly, the largest region of conserved co-expression was split in two modules of similar sizes. Whether this was accurate or not is hard to say, but will be discussed further in the light of functional importance in section 5.3.

The scale-free characteristics show that most genes in the data co-express strongly with only one or a few other genes, while a small number of genes are highly connected. Based on the assumption that the correlations are not random but caused by some underlying biological function, these network hubs are of special interest. In this work, the nodes with 20 or more co-expressed partners (hubs) are likely to play a special role in AD. Since these have strong associations with many other genes, they are more likely to be essential and might have regulatory roles [8]. Chowdhury et al. suggest that the genes which change behaviour across conditions with respect to a significant number of neighbors are the most interesting for biomarker identification [47]. It is important to note that the genes in the network are inferred because of their interactions, but we mostly focus on hubs rather than the interactions directly because it is less challenging and more annotated information is available in the literature.

5.1.2 Functional enrichment

Most of the enriched processes and pathways relate somehow to the progress of AD, and have been associated with the disease in previous studies. There was also an overall 2-fold enrichment of genes previously affiliated with AD in the CSD network. This verifies a certain quality of the inferred network in representing AD-specific gene correlations. At the same time, given the complexity of the disease, it would not be surprising to find

a variety of cellular functions enriched.

From the enrichment analyses we saw that there was an overall trend in significant biological processes primarily showing up in regions of conserved co-expression. This was the case both for whole networks and for the modules. Given the large number of S- and D-type links passing the significance threshold, it is unlikely that they were included randomly in the network. It could reflect a limitation in the use of GO annotation, being less sensitive to situations were a single gene has a drastic effect. GO is a quantitative rather than qualitative measure; each gene in the list counts only one time no matter how important the individual genes might be for that process. This could be a possible explanation for the lack of enrichment, especially in D-linked genes. It could be that the specific (S) and differentiated (D) types of co-expression patterns are involved in a variety of different processes that are not located closely in the network. Oppositely, in the regions of gene pairs that are strongly co-expressed in both conditions (C-type) there seems to be more genes working together in tandem, which could explain why there is an over-representation in biological processes.

5.2 Integration of DEA with CSD

The differential expression analysis (DEA) performed in this work resolved one of the limitations of the CSD method. Its focus on gene pair *correlations* might miss genes with significant changes in gene expression *levels* between conditions. Such DEGs do not necessarily have strong enough co-expression with other genes to be included in the CSD network. The integration of DEA with CSD allowed the simultaneous identification of differentially expressed genes (DEGs) and differentially co-expressed genes (DCGs) in one network, as visualized in Fig. 4.9.

Although the levels of differential mean expression is more commonly used as a threshold *before* generating co-expression networks, this approach was not chosen in this thesis. This was because we would lose information of gene pairs that show coordinated correlation patterns but are not individually DEGs, which was confirmed by the results. We saw that most of the genes in the CSD network were not differentially expressed (DEGs). This showed that changes in gene pair correlations can occur in the absence of individual differential expression, as has been found in previous studies as well [116]. The CSD framework hence captured changes in regulatory patterns that would not be detected by traditional DEA alone. Yet, the network *was* enriched with genes showing individual differential expression, in addition to their gene pair correlated expressions. This was particularly evident for the regions with conserved co-expression, which showed a 3-fold enrichment in DEGs. Another observation was that most of the DEGs in conserved region were closely connected, while the DEGs in the S/D-region were more individual and dispersed. This can be partly explained by the nature of the correlations, but the interpretation is not straightforward.

Interestingly, the DEA has thus provided a new perspective to the C-links of the CSD method by showing that they are not necessarily disease-independent. Although this should be validated, it demonstrates that the inclusion of C-type links in the method

is highly valuable. To my knowledge, and based on the method comparisons made by Voigt et al. [3], none of the other existing methods for differential co-expression integrates *conserved* interactions in the resulting network. Most disease studies using co-expression networks focus solely on the correlation *changes* and might therefore miss coordinately dysregulated genes as those found in this thesis.

5.3 Regions with conserved co-expression

The genes with homogeneously conserved type of co-expression were separated in two main regions in the CSD network (Fig 4.1), and further divided in three modules. The detection of enriched processes related to AD, combined with differentially expressed genes (DEGs) and previously AD-affiliated genes, make these modules and their hubs interesting for further studies.

The results indicate that module 4 and 6, especially, are potential disease modules. Most of the previously AD-related genes were found here (> 6-fold enrichment), indicating that these might actually be disease modules, despite the conserved co-expression. The DEA revealed that the *level* of gene activity was changed between the two conditions even though the *correlation patterns* between the genes were conserved. One hypothesis could be that strong correlations are maintained because the genes are tightly co-regulated or form part of complexes that are collectively altered in the diseased. As expected, all closely connected DEGs were either up-regulated *or* down-regulated. When one gene falls out of control, its co-expressed neighbor does so too due to strong positive correlation. Specifically, in the largest conserved region, all of the DEGs - including the highly connected hubs - were down-regulated to some extent. This means that the processes and pathways found enriched in these two modules are likely to be down-regulated in AD patients compared to healthy controls. Genes that are down-regulated in AD might represent dysregulations important for disease progress. In fact, most of the significantly enriched terms have been associated with AD development. This includes important processes involved in signaling, synaptic activity, regulation of neurotransmitter levels, intracellular protein transport and oxidative phosphorylation. The results from GO enrichment of all down-DEGs also indicated that AD brains show a loss in the maintenance of synapse structure. The disruption of various aspects of synaptic function corresponds well with the established role of synaptic loss in AD patients, which eventually leads to memory impairment [22, 14].

The results of the enrichment analyses and the functional annotation of the two hubs of module 4 indicate that the module is involved in processes related to neurotransmitter release. In particular, high enrichment was found in KEGG pathways related to GABAergic synapse. GABA is the main inhibitory neurotransmitter in the CNS, and a reduction in the levels of this neurotransmitter has been observed in AD patients [117]. The transport and release of neurotransmitters happens with the aid of synaptic vesicles, explaining why this module was also enriched in processes such as vesicle fusion and docking (Table 4.4 and the "synaptic vesicle cycle" in general. Several genes in the module form part of the SNARE complex, essential for the mentioned processes, which includes the hub NAPB and its PPI-associated partner SYT1. The SNARE com-

plex, and neurotransmitter release in general, is dependent on calcium. Impaired Ca^{2+}-homeostasis has been shown to follow from toxic Aβ-generation. VSNL1 is a calcium sensor protein, and its misregulation has been associated with impaired neuroprotection, synaptic plasticity and eventually cell death [118].

Module 6 had the largest number of hubs represented, and was enriched in numerous pathways that have earlier been associated with AD. Several pathways related to PI3K-Akt signaling, which is an important intracellular signal transduction pathway that regulates the cell cycle [119]. The over-represented pathways include VEGF- and mTOR signaling (Table 4.6). The Vascular Endothelial Growth Factor (VEGF) is a diverse growth factor that can stimulate both the PI3K-Akt and MAPK pathway [4]. Aberrant VEGF signaling has been linked to neurodegeneration through the disruption of the blood-brain barrier [120]. The mammalian target of rapamycin (mTOR) is part of protein complexes downstream of the PI3K-Akt signaling pathway [4]. It is a serine/threonine protein kinase with a key role in the negative regulation of autophagy, a pathway involved in the degradation of abnormal proteins [4, 121]. This corresponds well with the "regulation of macroautophagy" found in GO analysis. Autophagy malfunction has been associated with AD, where it influences the generation and metabolism of $A\beta$ [121]. The enrichment in "cellular senescence" (Table 4.6) also indicates that there might be an interruption of the cell cycle progress. Many of the enriched genes in the pathways of this module were previously AD-related genes, which supports that this kind of signaling is important for the disease progress.

Module 9 showed quite different trends compared to the other two C-modules, which fits with the distance seen in the network. All DEGs identified in this isolated C-region were *up-regulated* in AD, though only a few were significant (Table 4.9). It is therefore difficult to determine the biological role of the genes within this module without further detailed analysis. There were less terms enriched in this module, which might be due to the lower gene set size (N = 66). Still, those that *were* enriched are processes mainly involved in neuron development, such as myelination. This includes metabolism of sphingolipids, key lipids in the protective myelin layers of neurons [88]. To the best of my knowledge, not much is known about the role of myelin in AD, but myelin damage has been observed [122]. The potential upregulation found in this work might be a compensatory response to the deleterious mechanisms of AD, but more research is needed.

By comparing the functional enrichment found for modules 4 and 6 individually, it was possible to hint at whether the modular partitioning was biologically reasonable or not. From a first glance at the GO results (Table 4.4) it looked like the modules represented quite different biological processes. This would confirm the quality of the modular partitioning. With a more thorough comparison of all GO terms, several processes were found in common between the two modules, some more specific than others. This could explain why the two modules are closely interconnected in the network. The most specific terms in common were "chemical synaptic transmission", "regulation of synaptic vesicle cycle" and "regulation of exocytosis" (appendix B.2). These are biological processes related to signaling and transport that are essential for the information flow within and between nerve cells. Given the established role of aberrant signaling in

AD, it is supportive to find such processes coordinately down-regulated [18].

These two modules of the largest conserved region could collectively be regulated by some of the "bridge" nodes found in-between the two unique parts of the giant component. With their high betweenness centrality they represent connectors between the region of conserved co-expression and differential co-expression (S&D). Some of these genes *did* show changed correlation patterns in AD patients, and their prominent position make them important candidates with regulatory roles. One of these was the AD-affiliated gene *GSK3B*, well-known for its role in pathological role in AD [106]. Its "bridge"-location in the network further supports its importance in signal transduction. Glycogen synthase kinase-3 (GSK3) is also sometimes called "tau kinase I" due to its role in phosphorylating the protein known to form the abnormal tau protein [106]. Directly connected in the network was another potential disease-genes based on guilt by association, namely *GRSF1*. In addition to its S-link to GSK3B, it is connected to three other AD-related genes and has S/D-type co-expression with all its neighbors. It encodes an RNA-binding protein (RBP), and such proteins have been associated with neurological disorders due to their regulation of local mRNA translation at the synapses [108, 123]. Finally, SMYD2 has an interesting location, connecting VSNL1 of module 4 to the highest degree node (KIAA1841) in the network. *SMYD2* encodes a protein methyltransferase, which by methylating other proteins, such as the tumor suppressor p53, functions as a regulator of cell proliferation and cancer [104]. Yi et al. has stated that "Methylation on substrates always cross-talks with other posttranslational modifications, especially phosphorylation, to affect signaling pathways and target genes related to cancer and other disease." [104]. In this way, it might be involved in the regulation of signaling pathways related to the development of AD.

5.4 Region with specific and differentiated co-expression

The S&D-region seems to be governed mainly by individual genes with influential roles rather than collective trends (observed in C-regions). Although the S-type and D-type networks had a larger gene set for the enrichment analyses, the result was still fewer significant terms with lower enrichment values. This made it somewhat challenging to infer global biological interpretations. Nevertheless, the region included some of the largest hubs, and their change in pairwise correlations with many neighbors make them interesting for further studies. Most of these hubs have "pure" differential co-expression, co-expression not associated with a change in mean expression levels, making them rewiring candidates [116]. The details of the changes in correlation values from healthy to sick would be interesting to inspect further. All hubs were examined considerably in chapter 4; hence only some will be discussed here, focusing on aspects highly relevant for AD pathology. Genes in the S-network were slightly enriched in a few broad categories, involving protein localization, transport, and signaling. A more detailed analysis of prominent genes highlighted signaling and lipid transport necessary for APP processing, which will be discussed in the following paragraphs.

Emerging evidence suggests that impaired energy metabolism is characteristic of AD brains and that disrupted insulin and glucose metabolism can increase the risk of de-

veloping AD [110, 109]. In particular, the Hisayama study found that insulin resistance is associated with an increased risk of senile plaques in the brain [124]. Although no significant KEGG enrichment was found within the S&D-region, a certain role in insulin/glucagon signaling was inferred. *PHKG1* was prominent, having specific (S) co-expression with two of the largest hubs, KIAA1841 and TMEM178A. It encodes a subunit of a serine/threonine-protein kinase and might have an important regulatory role [78]. This is also supported by its S-link to *ADAMTS2*, the most up-regulated gene in the CSD network. Members of the ADAM-family have been identified with α-secretase-activity, cleaving APP within the Aβ sequence and blocking pathological cleavage into amyloid peptides [125]. *ADAMTS2* might thus have a neuroprotective role, but this needs further investigation. Defective insulin signaling was also supported by the involvement of the previously AD-related gene *IRS1*, which was S-linked to the hub PLTP. In total, these results indicate that changes in the regulatory patterns of this pathway play important roles in AD.

Several of the hubs encode proteins related to lipid metabolism and transport. This includes PLTP, which had a central position in the network, connecting many of the other hubs of the S&D-region. These hubs might regulate or form part of cell-signaling platforms, called *lipid rafts*, where pathogenic signaling that underlie the neuropathology in AD could happen [126]. The formation of such lipid rafts, which are cholesterol- and sphingolipid-rich microdomains within the cell membrane, has been implicated in several neurological diseases [126]. Past research has reported that Phospholipid Transfer Protein (PLTP) has a role in the processing of APP into Aβ, which is tightly linked to both lipid homeostasis and AD progression. In particular, Mansuy et al. proposed that elevated activity of *PLTP* leads to a higher amount of Aβ in the brain [127]. In the DEA performed in this thesis, *PLTP* was slightly up-regulated, but not significantly ($\log_2$FC = 0.23, q = 0.11). Although dysregulation of lipid homeostasis is linked to AD, it is unclear whether the altered lipid levels are the cause or consequence of AD [21].

The most highly connected node in the region, and network as a whole, was a novel transcript with minimal functional annotation; KIAA1841. The GO enrichment in section 4.1.6 did not lead to any biological insight into module 26, of which KIAA1841 contains a large proportion of connections. Interestingly though, one KEGG pathway was significantly enriched in the module, namely the Notch signaling pathway (see section 4.1.6). This pathway has been associated with neurodegeneration and is partly regulated by the infamous AD-gene *PSEN1* [128]. This disease-gene was located in the network, but with a modest role outside of the modules ("Other" in table 4.7). Yet, another AD-related gene with a similar role was directly connected to KIAA1841, namely *APH1B*. This gene codes for a subunit of the γ-secretase complex, which can cleave both Notch receptors and APP [78]. This could explain why Notch signaling was enriched, and further indicate that KIAA1841 has a central role in the disease. The more extensive literature search needed to conclude the biological relevance of all these genes is left for future work due to time constraints.

5.5 Relation to PPIs

An important limitation to correlation networks is that no causal interpretation of a link between gene pairs can be inferred directly. However, several studies have shown that subunits of a protein complex or otherwise interacting proteins tend to show similar patterns of gene expression [56, 58]. The motivation of integrating PPI was therefore to investigate the potential physical interaction between the protein products of co-expressed genes. Surprisingly, only three pairs of gene interactions were common to both the PPI- and CSD network. Two of these were C-type links in the CSD network, while the other was a D-link. The fact that two genes show a high degree of gene expression correlation does not seem to correspond with the existence of a protein-protein interaction. The same was found in a previous study comparing co-expression and PPI networks in yeast [129]. However, the three interactions found were interesting from a disease perspective, as explored in section 4.2.2, and should be investigated further.

The PPI data was chosen because of its unbiased high quality, but it might have been too stringent for comparison with the correlation network. In addition, only direct interactions from the *binary* interaction map was investigated, hence potentially underlying indirect biological relationships might have been missed by this approach. The limitations of Y2H might explain some of the lack of identified protein interactions for the co-expressed genes. PPIs in the human interactome can remain undetected by screening for reasons such as post-translational processing in humans that does not occur in yeast, or transient binding stability of proteins [65]. It is important to remember that the process of gene expression is complex - much can happen from gene to protein.

5.6 Method and study limitations

Overall, the methods used in this thesis were successful in identifying transcriptomic changes in human AD brains. Nonetheless, various challenges and limitations were experienced and will be addressed in the following paragraphs.

First of all, there are several challenges associated with microarray data, and transcriptomic technologies in general. As experienced in this study, the original raw data requires several steps of processing and quality assessment. This includes the non-trivial task of annotating probes to the correct transcript (read: gene). It is also important to note that the resulting values are relative and not absolute measures. Relative measures allow for comparing change in gene expression, as was done in the DEA, but interpretations are not straightforward. The mRNA levels measured, being indicative of active genes, do not necessarily correlate with protein levels. To validate the microarray results, one could perform experimental procedures such as quantitative PCR (qPCR) and/or Western blot on the most interesting candidate genes. This however brings us to the challenge in deciding which genes to prioritize. Usually the ones with largest expression changes between the two conditions studied are chosen, but the threshold for this is arbitrary. Another important aspect of microarray is that each chip gives a *snapshot* of gene expression, and it reflects the *average* values of expressions across millions of cells (from a single experiment). The average may not be a good representation of the

distribution of gene expression levels, which may vary between different cell types of the brain. Latest advances in RNA-seq have led to single cell-RNA seq, which addresses this issue [47]. It allows researchers to study cell-type-specific transcriptional changes, which could be used to incorporate the cellular heterogeneity of the brain. Overall, the use of microarray data was considered sufficient for the aim of this thesis. It could however be interesting to compare with gene expression data from RNA-seq, which are able to quantify more non-coding RNAs and might give an even better understanding of the regulatory mechanisms of the disease.

A downstream analysis is highly dependent on the quality of the original data [8]. It was assumed that the data accessed from the public gene expression database should be of high-quality and that poor results were more likely to result from improper sample selection. However, the use of only one individual data set can raise study-specific bias. Although the sample count from the study used was high, which has been found highly important by many previous studies [130, 131], there could be confounding factors not accounted for. For instance, some of the different samples extracted were taken from different brain regions within the same individual. This is violating the assumption of independent observations which is made when calculating correlation. For these reasons, the results should be validated by at least one other independent data set. Additionally, the assumption made for excluding variance calculation was not verified in this work and might therefore have influenced the results.

Although it is generally recommended to use a large sample size, there are potential downsides of pooling together multiple samples. In this thesis, combining all four brain regions could explain the lack of specificity in enrichment results and the little change in mean expression levels observed from DEA. In agreement, the DEA performed on individual regions *did* show larger $\log_2$FC-values (appendix G). Although AD is generally thought to affect most brain regions eventually, some might not be as affected by the disease. For example, the postcentral gyrus - containing the primary somatosensory cortex - has been hypothesized to be relatively unaffected by AD [69]. This was also seen from the lack of significant DEGs in this region (appendix G). However, more recent studies have found functional changes in this area associated with mental disorders such as major depression and Schizophrenia [132, 133]. Since this area of the brain processes information such as touch, pain, and emotion [133] - it could be partly responsible for mood changes observed in AD patients. Therefore, a more detailed study into each of the four regions could be interesting, especially the hippocampus, which had significant DEGs (appendix G). Regardless, the work performed in this thesis should capture unique AD signatures across all brain regions compared to non-AD subjects.

The DEA was quite easily integrated with the established CSD framework and is a general approach that can be used for any pairwise comparisons in future systems biology research. Still, some challenges were experienced in this thesis that should also be addressed before future employment. The calculations of change in transcript abundance for all genes were performed by "manual" scripts made in Python. Further use of DEA as part of the CSD framework could benefit from using already established packages, such as *limma* in R [134]. The most challenging aspect was the subsequent establish-

ment of an appropriate threshold for DEG identification. Adjusting for multiple testing is necessary, and this was done with the BH-method. The result was a list of genes that were more likely to be indeed differentially expressed, with a false discovery rate of 5%. As expected, the volcano plot showed a general trend of increasing significance with increasing fold change. However, numerous genes were considered significant based on the multiple t-tests but had a low (almost no) magnitude of change. These were filtered out before integration with CSD. Whether the chosen limit of $logFC = \pm 0.2$ was an appropriate threshold or not is debatable. This factor influences all downstream predictions and should thus be carefully assessed. Given that FC is a relative measure, it was decided valid to define DEGs as the genes with the greatest change compared to the distribution of all genes in the data. The combination with significance in a "double"-filtration approach highly increased the liability.

Modules were predicted solely from the wiring diagram of the CSD network and identified whether they are actually there or not. It is important to note that these community structures do not necessarily reflect "real" communities. The Louvain community detection algorithm used forms non-overlapping communities by assigning all genes to only one distinct community, even though some genes might not even belong to a community, or might be part of several. From a biological perspective, community overlaps would make sense as several genes are often involved in multiple biological processes and pathways. This could be a reason for the lack of over-representation of biological processes in the S- and D-type modules, which seem to "overlap" quite a bit in the network. At the same time, there is no guarantee for clusters of co-expressed genes being co-regulated or part of the same biological process [47].

Enrichment analyses are associated with various challenges, both within science in general and for this work in particular. The main factors to consider are the input size and reference background. Enrichment scores are dependent on gene set size, and since the lists used in this work are not of equal length, it might not be valid to compare these directly. Regarding the reference list, the total genes expressed in the microarray data could have been used rather than the default database. Also, using the same database for both GO and Pathways, e.g. PANTHER, could lead to easier comparisons. Another limitation in the GO annotation process is establishing a balance between too broad and too specific terms. Broader categories include a higher number of genes, but are not necessarily specific enough to give more insight. In this work, the functional processes and pathways with the largest FE were often caused by only a few genes. These are over-represented compared to what would be expected by chance, but represent only a small proportion of the network. Given the small set of genes enriched, it can be challenging to draw conclusions for larger parts of the network. It is also important to note that enrichment analysis is based on *statistical* significance and does not directly imply the underlying biological event(s) [111]. For these reasons, enrichment analyses should only be used as a guideline for biological annotation.

The implementation of the CSD method showed a high performance, but the software used still has room for improvement. The time it took to calculate all pairwise gene correlations was long, given the large data set, even after excluding variance calculation. Being the rate-limiting step, further reducing the time complexity here would be very

useful. Further work with the CSD framework would also largely benefit from combining the three software programs into one, possibly making a Graphical User Interface (GUI) for more user-friendly implementation and easier manipulation of parameters.

Chapter 6

Conclusion & Outlook

This thesis aimed to identify transcriptomic changes specific to AD by using an integrative approach based on the CSD framework. Published gene expression data from human brain tissue was used to compare expression profiles in AD patients to age-matched cognitively normal controls. Together, the work done in this thesis shows the advantage of using systems- and network biology approaches in complex diseases such as AD.

The constructed differential co-expression network successfully captured gene expression patterns that indicate response mechanisms to the change from healthy to diseased. As much as 64 previously AD-affiliated genes were recognized in the network, which is both indicative of high-quality network inference and at the same time allows further biological predictions of pathogenesis. Potential roles of neighboring genes were inferred based on guilt by association. The CSD network revealed many prominent genes and pathways likely involved in AD development. Interestingly, novel genes with minimal functional annotation were identified, such as the network hub KIAA1841. Being the most highly connected node in the network, it is quite possibly essential. The hub changes behavior with respect to all its neighbors across the two conditions and might thus have a regulatory role in AD. Based on modular enrichment and the functions of its neighbors, it might be involved in signaling pathways related to lipid- and Aβ metabolism. Hence, KIAA1841 is a new candidate for a role in neurodegenerative diseases such as AD, and should be investigated further.

The most significant contribution to the method framework was the differential expression analysis (DEA) performed in this work. The two main methods - differential expression and differential *co*-expression - complemented each other, each providing information that the other did not capture. The CSD method identified many genes with pairwise correlation changes that did not show individual changes in mean expression. Oppositely, the DEA found DEGs both outside and within the network, offering new insights not captured by the differential co-expression network alone. 229 nodes from the

original 1535 nodes in the CSD network were recognized as DEGs. These were enriched in the regions of conserved (C) co-expression, where there was a trend of similarly regulated genes connecting with each other, suggesting a coordinated regulation. In the largest conserved (blue) region of the network, including modules 4 and 6, all differentially expressed genes (DEGs) were down-regulated in AD. Hence, although gene pair correlations were maintained (strong positive correlations across samples in both conditions), the mean expression levels decreased from control to sick.

Based on the high enrichment in disease association, combined with down-regulated genes enriched in AD-associated processes, modules 4 and 6 are promising disease module candidates. Both GO and KEGG analyses pointed out that the genes of this module are involved in the synaptic vesicle cycle and might be responsible for impaired signal transmission. This is a common pathological feature of AD and other neurodegenerative diseases [19]. Interestingly, all the C-type hubs had at least one AD-related neighbor, making these hubs disease gene candidates. This includes genes such as *YWHAH*, which has earlier been associated with Schizophrenia and was in this network connected to four AD-related genes. The decrease in the expression of this transcription activator might explain some of the down-regulated processes.

This work has, through DEA, demonstrated the importance of including the conserved type of co-expression and not only the differential types of co-expression when comparing gene expression profiles. This is something that the CSD method does that is not generally seen in other studies of co-expression networks [3]. The genes with C-type co-expression from the CSD framework have usually not been considered valuable from a disease perspective. The DEA put a new perspective on this by showing that the *level* of gene expression was affected by disease even if the pairwise correlations were conserved. Although salient, this finding should be validated by implementing the same method on an independent data set. Moreover, it would be highly interesting to study the correlation between differential expression and differential co-expression. Besides, filtering based on mean expression levels *before* network generation could be considered for a more strictly defined CSD network, potentially highlighting the conserved areas even more.

More detailed analyses of the relationships behind the correlation patterns are imperative to understand their biological relevance to AD better. Experimental validations of annotated gene functions and the inferred associations would be a good starting point. Based on the PPI integration, there was no strong relation between correlated gene expression and previously identified physical connections between the encoded proteins. Only three gene interactions in the CSD network had known protein associations. For further studies, it could be interesting to investigate more indirect PPI links, which might explain more of the relationships behind the CSD links. A more extensive database with lower-quality PPIs, such as BIOGRID or STRING, could also be considered, but the validity of predictions would need careful assessment. Additional types of data could be integrated for more details into the underlying mechanisms of AD pathogenesis. For instance, it would be interesting to incorporate public TF data to explore the potential regulatory roles of the genes involved in the identified signaling pathways.

Alzheimer's disease is heterogeneous and complex, so it was not surprising to find many different biological pathways enriched in the network. Only some of them were explored in this thesis and it cannot be ruled out that other pathways may be more essential in the disease progression. The same applies to the discussion of previously AD-affiliated genes, and thus future studies could investigate the roles of each of the 64 genes and their immediate neighborhoods. It could also be interesting to compare with AD-associated genes from other databases such as KEGG or DisGeNET [4, 135]. Further, it could be of interest to apply the CSD framework on gene expression profiles from other neurodegenerative diseases, such as Parkinson's disease, that were enriched in the network. The resulting CSD networks could be compared to look for more disease-specific expression patterns.

Although the work performed in this thesis answers the aim of capturing universal signatures of AD pathogenesis, other aspects of the microarray data set could be addressed in further research. First of all, given that aging is a major risk factor for AD, future studies could estimate this potential confounding factor. Even though this work controlled confounding by design, further work could employ a multivariate statistical analysis. Then, besides comparing with the network from the complete data set (w/ young controls), it could be even better to compare with a CSD network created solely from controls. This network would identify genes specific for aging by comparing young vs. old. Only the diseased as a whole group (n = 80) was used in this thesis. Future work could get a more detailed view by including Braak stages and APOE genotype, both of which were registered in the microarray experiment [67] but omitted here for simplicity. Also, to account for tissue-specificity, four smaller CSD networks could be generated for each region (HC, EC, SFG and PCG) from the AD and healthy individuals of >= 60 years. In addition, it would be interesting to check if some specific cell types were enriched in the network, for example by using Enrichr [51, 83]. If so, cell heterogeneity could be considered by implementing CSD on single-cell transcriptomic data from major brain cell types in patients with AD. Regardless, given the limitations of the applied data set and the method in general, one of the first steps should be to compare the results with validated reference data. Preferably, a meta-analysis can be used to find universal patterns across data sets that give stronger predictions than those made using individual data sets [58].

All in all, a combination of several approaches, also involving the verification of functional annotation by more detailed experiments, is needed to understand the complex mechanisms underlying the pathogenesis of AD. The ultimate goal would be to translate the knowledge obtained from this research into treatments that prevent, slow down, or cure AD. Finally, beyond answering the aim of this thesis, the integrated approach provides a powerful platform that can be useful for future comparative studies in systems biology.

Bibliography

[1] Geneva: World Health Organization; 2017, ed. *Global action plan on the public health response to dementia 2017–2025*. Licence: CC BY-NC-SA 3.0 IGO. ISBN: 978-92-4-151348-7.

[2] Krishnarao Appasani, ed. *Bioarrays - From Basics to Diagnostics*. Humana Press, 2007. ISBN: 978-1-58829-476-0.

[3] André Voigt, Katja Nowick, and Eivind Almaas. "A composite network of conserved and tissue specific gene interactions reveals possible genetic interactions in glioma". In: *PLOS Computational Biology* 13.9 (Sept. 2017). Ed. by Lilia M. Iakoucheva, e1005739. DOI: 10.1371/journal.pcbi.1005739.

[4] M. Kanehisa. "KEGG: Kyoto Encyclopedia of Genes and Genomes". In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 27–30. DOI: 10.1093/nar/28.1.27.

[5] Hao Chi, Hui-Yun Chang, and Tzu-Kang Sang. "Neuronal Cell Death Mechanisms in Major Neurodegenerative Diseases". In: *International Journal of Molecular Sciences* 19.10 (Oct. 2018), p. 3082. DOI: 10.3390/ijms19103082.

[6] Srdjan Kesić. "Systems biology, emergence and antireductionism". In: *Saudi Journal of Biological Sciences* 23.5 (Sept. 2016), pp. 584–591. DOI: 10.1016/j.sjbs.2015.06.015.

[7] A-L Barab´asi, ed. *Network Science*. Cambridge University Press, 2015. ISBN: 978-1107076266.

[8] Bjrn H. Junker and Falk Schreiber, eds. *Analysis of Biological Networks*. John Wiley & Sons, Inc., Feb. 2008. DOI: 10.1002/9780470253489.

[9] Alberto de la Fuente. "From 'differential expression' to 'differential networking' – identification of dysfunctional regulatory networks in diseases". In: *Trends in Genetics* 26.7 (July 2010), pp. 326–333. DOI: 10.1016/j.tig.2010.05.001.

[10] Clark David P. and Nanette J. Pazdernik, eds. *Molecular Biology. 2nd ed.* Academic Press, Elsevier, 2013. ISBN: 978-0123785947.

[11] Thomas J. Gonda and Robert G. Ramsay. “Directly targeting transcriptional dysregulation in cancer”. In: *Nature Reviews Cancer* 15.11 (Oct. 2015), pp. 686–694. DOI: 10.1038/nrc4018.

[12] Wei Jin et al. “Dysregulation of Transcription Factors: A Key Culprit Behind Neurodegenerative Disorders”. In: *The Neuroscientist* 25.6 (Nov. 2018), pp. 548–565. DOI: 10.1177/1073858418811787.

[13] Theo Vos et. al. “Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015”. In: *The Lancet* 388.10053 (Oct. 2016), pp. 1545–1602. DOI: 10.1016/s0140-6736(16)31678-6.

[14] D. M. Holtzman, J. C. Morris, and A. M. Goate. “Alzheimer’s Disease: The Challenge of the Second Century”. In: *Science Translational Medicine* 3.77 (Apr. 2011), 77sr1–77sr1. DOI: 10.1126/scitranslmed.3002369.

[15] Lisette J A Kogelman et al. “Identification of co-expression gene networks, regulatory genes and pathways for obesity based on adipose tissue RNA Sequencing in a porcine model”. In: *BMC Medical Genomics* 7.1 (Sept. 2014). DOI: 10.1186/1755-8794-7-57.

[16] Marta R. Moksnes. “Identification of Novel Genes associated with Rheumatoid Arthritis using Differential Gene Co-Expression Analysis”. MA thesis. NTNU Trondheim, May 2018.

[17] M. Gulla. “An integrated systems biology approach to investigate transcriptomic data of thyroid carcinoma”. MA thesis. NTNU Trondheim, May 2019.

[18] Eliza Courtney et al. “Transcriptome profiling in neurodegenerative disease”. In: *Journal of Neuroscience Methods* 193.2 (Nov. 2010), pp. 189–202. DOI: 10.1016/j.jneumeth.2010.08.018.

[19] Michael A. DeTure and Dennis W. Dickson. “The neuropathological diagnosis of Alzheimer’s disease”. In: *Molecular Neurodegeneration* 14.1 (Aug. 2019). DOI: 10.1186/s13024-019-0333-5.

[20] Stephen Todd et al. “Survival in dementia and predictors of mortality: a review”. In: *International Journal of Geriatric Psychiatry* (Mar. 2013). DOI: 10.1002/gps.3946.

[21] Marcus O. W. Grimm, Tatjana L. Rothhaar, and Tobias Hartmann. “The role of APP proteolytic processing in lipid metabolism”. In: *Experimental Brain Research* 217.3-4 (Dec. 2011), pp. 365–375. DOI: 10.1007/s00221-011-2975-6.

[22] M. Sheng, B. L. Sabatini, and T. C. Sudhof. “Synapses and Alzheimer’s Disease”. In: *Cold Spring Harbor Perspectives in Biology* 4.5 (Apr. 2012), a005777–a005777. DOI: 10.1101/cshperspect.a005777.

[23] “Memory loss in Alzheimer’s disease”. In: *Memory* 15.4 (Dec. 2013), pp. 445–454. DOI: 10.31887/dcns.2013.15.4/hjahn.

[24] H. Braak and E. Braak. “Neuropathological staging of Alzheimer-related changes”. In: *Acta Neuropathologica* 82.4 (Sept. 1991), pp. 239–259. DOI: 10.1007/bf00308809.

[25] Noa Rappaport et al. “MalaCards: an integrated compendium for diseases and their annotation”. In: *Database* 2013 (Jan. 2013). DOI: 10.1093/database/bat018.

[26] S. Glenna. *Figshare. Master's project in Systems biology; Transcriptomic study of Alzheimer's disease.* 2020.

[27] Bruno Dubois et al. “Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria”. In: *Alzheimer's & Dementia* 12.3 (Mar. 2016), pp. 292–323. DOI: 10.1016/j.jalz.2016.02.002.

[28] Bushra Imtiaz et al. “Future directions in Alzheimer's disease from risk factors to prevention”. In: *Biochemical Pharmacology* 88.4 (Apr. 2014), pp. 661–670. DOI: 10.1016/j.bcp.2014.01.003.

[29] Kay Deckers et al. “Target risk factors for dementia prevention: a systematic review and Delphi consensus study on the evidence from observational studies”. In: *International Journal of Geriatric Psychiatry* 30.3 (Dec. 2014), pp. 234–246. DOI: 10.1002/gps.4245.

[30] Matt Paradise, Claudia Cooper, and Gill Livingston. “Systematic review of the effect of education on survival in Alzheimer's disease”. In: *International Psychogeriatrics* 21.01 (Nov. 2008), p. 25. DOI: 10.1017/s1041610208008053.

[31] Yaakov Stern. “Cognitive reserve in ageing and Alzheimer's disease”. In: *The Lancet Neurology* 11 (Nov. 2012), pp. 1006–1012. DOI: 10.1016/s1474-4422(12)70191-6.

[32] B. Kolb and R. Gibb. “Brain plasticity and behaviour in the developing brain.” In: *Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie canadienne de psychiatrie de l'enfant et de l'adolescent* 20(4) (2011), pp. 265–76.

[33] Clive Ballard et al. “Alzheimer's disease”. In: *The Lancet* 377.9770 (Mar. 2011), pp. 1019–1031. DOI: 10.1016/s0140-6736(10)61349-9.

[34] Yvonne S. Eisele et al. “Targeting protein aggregation for the treatment of degenerative diseases”. In: *Nature Reviews Drug Discovery* 14.11 (Sept. 2015), pp. 759–780. DOI: 10.1038/nrd4593.

[35] Béla Bollobás. *Modern Graph Theory*. Springer New York, 1998. DOI: 10.1007/978-1-4612-0619-4.

[36] ColiN00B. 2020. URL: https://www.needpix.com/photo/911619/nerve-cell-neuron-brain-neurons-nervous-system-synapse-cells-think-neural-pathways (visited on 11/24/2020).

[37] Enrico Pieroni et al. “Protein networking: insights into global functional organization of proteomes”. In: *PROTEOMICS* 8.4 (Feb. 2008), pp. 799–816. DOI: 10.1002/pmic.200700767.

[38] Eberhard O. Voit, ed. *A First Course in Systems Biology*. Garland Science, 2017. ISBN: 9780815345688.

[39] Réka Albert and Albert-László Barabási. “Statistical mechanics of complex networks”. In: *Rev. Mod. Phys.* 74 (Jan. 2002), pp. 47–97. DOI: 10.1103/RevModPhys.74.47.

[40] Anna D. Broido and Aaron Clauset. “Scale-free networks are rare”. In: *Nature Communications* 10.1 (Mar. 2019). DOI: 10.1038/s41467-019-08746-5.

[41] M. E. J. Newman. “Mixing patterns in networks”. In: *Physical Review E* 67.2 (Feb. 2003). DOI: 10.1103/physreve.67.026126.

[42] Albert-László Barabási and Zoltán N. Oltvai. “Network biology: understanding the cell’s functional organization”. In: *Nature Reviews Genetics* 5.2 (Feb. 2004), pp. 101–113. DOI: 10.1038/nrg1272.

[43] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. “Network medicine: a network-based approach to human disease”. In: *Nature Reviews Genetics* 12.1 (Dec. 2010), pp. 56–68. DOI: 10.1038/nrg2918.

[44] Daniel A. Schult Aric A. Hagberg and Pieter J. Swart. “Exploring network structure, dynamics, and function using NetworkX”. In: *Proceedings of the 7th Python in Science Conference (SciPy2008)* (Aug. 2008), pp. 11–15.

[45] Vincent D Blondel et al. “Fast unfolding of communities in large networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (Oct. 2008), P10008. DOI: 10.1088/1742-5468/2008/10/p10008.

[46] M. J Snustad D. P. Simmons, ed. *Principles of Genetics*. Singapore: John Wiley & Sons, Inc., 2012.

[47] Hussain Ahmed Chowdhury, Dhruba Kumar Bhattacharyya, and Jugal K. Kalita. “(Differential) Co-Expression Analysis of Gene Expression: A Survey of Best Practices”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2019), pp. 1–1. DOI: 10.1109/tcbb.2019.2893170.

[48] Mohan S. Rao et al. “Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies”. In: *Frontiers in Genetics* 9 (Jan. 2019). DOI: 10.3389/fgene.2018.00636.

[49] A Oberthuer et al. “Comparison of performance of one-color and two-color gene-expression analyses in predicting clinical endpoints of neuroblastoma patients”. In: *The Pharmacogenomics Journal* 10.4 (July 2010), pp. 258–266. DOI: 10.1038/tpj.2010.53.

[50] John Quackenbush. “Microarray data normalization and transformation”. In: *Nature Genetics* 32.S4 (Dec. 2002), pp. 496–501. DOI: 10.1038/ng1032.

[51] Edward Y Chen et al. “Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool”. In: *BMC Bioinformatics* 14.1 (2013), p. 128. DOI: 10.1186/1471-2105-14-128.

[52] Xiangqin Cui and Gary A Churchill. “Statistical tests for differential expression in cDNA microarray experiments”. In: *Genome Biology* 4.4 (2003), p. 210. DOI: 10.1186/gb-2003-4-4-210.

[53] Gunnar G. Løvås, ed. *Statistikk for universiteter og høgskoler*. Universitetsforlaget, 2004. ISBN: 9788215002248.

[54] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1 (1995), pp. 289–300.

[55] Elise A. R. Serin et al. "Learning from Co-expression Networks: Possibilities and Challenges". In: *Frontiers in Plant Science* 7 (Apr. 2016). DOI: 10.3389/fpls.2016.00444.

[56] M. B. Eisen et al. "Cluster analysis and display of genome-wide expression patterns". In: *Proceedings of the National Academy of Sciences* 95.25 (Dec. 1998), pp. 14863–14868. DOI: 10.1073/pnas.95.25.14863.

[57] M. P. S. Brown et al. "Knowledge-based analysis of microarray gene expression data by using support vector machines". In: *Proceedings of the National Academy of Sciences* 97.1 (Jan. 2000), pp. 262–267. DOI: 10.1073/pnas.97.1.262.

[58] Matthew T. Weirauch. "Gene Coexpression Networks for the Analysis of DNA Microarray Data". In: *Applied Statistics for Network Biology*. Wiley-VCH Verlag GmbH & Co. KGaA, Apr. 2011, pp. 215–250. DOI: 10.1002/9783527638079.ch11.

[59] Cecily J Wolfe, Isaac S Kohane, and Atul J Butte. In: *BMC Bioinformatics* 6.1 (2005), p. 227. DOI: 10.1186/1471-2105-6-227.

[60] Sipko van Dam et al. "Gene co-expression analysis for functional classification and gene–disease predictions". In: *Briefings in Bioinformatics* (Jan. 2017), bbw139. DOI: 10.1093/bib/bbw139.

[61] Gang-Guo Li and Zheng-Zhi Wang. "Evaluation of similarity measures for gene expression data and their correspondent combined measures". In: *Interdisciplinary Sciences: Computational Life Sciences* 1.1 (Mar. 2009), pp. 72–80. DOI: 10.1007/s12539-008-0005-3.

[62] Lin Song, Peter Langfelder, and Steve Horvath. "Comparison of co-expression measures: mutual information, correlation, and model based indices". In: *BMC Bioinformatics* 13.1 (2012), p. 328. DOI: 10.1186/1471-2105-13-328.

[63] A. Ardeshir Goshtasby. "Similarity and Dissimilarity Measures". In: *Image Registration*. Springer London, 2012, pp. 7–66. DOI: 10.1007/978-1-4471-2458-0_2.

[64] Michael E. Cusick et al. "Interactome: gateway into systems biology". In: *Human Molecular Genetics* 14 (Oct. 2005), R171–R181. DOI: 10.1093/hmg/ddi335.

[65] Javier De Las Rivas and Celia Fontanillo. "Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks". In: *PLoS Computational Biology* 6.6 (June 2010). Ed. by Fran Lewitter, e1000807. DOI: 10.1371/journal.pcbi.1000807.

[66] S. Glenna. *Github. Master's project in Systems biology; Transcriptomic study of Alzheimer's disease.* 2020. URL: https://github.com/susag3/Masters-project-in-Systems-biology-Transcriptomic-study-of-Alzheimer-s-disease.git.

[67] Berchtold NC. *Microarray analysis of Alzheimer's disease patients across 4 brain regions.* 2017. URL: www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-48350/.

[68] Awais Athar et al. "ArrayExpress update – from bulk to single-cell expression data". In: *Nucleic Acids Research* 47,Issue D1 (Oct. 2018), pp. D711–D715. DOI: 10.1093/nar/gky964.

[69] Nicole C. Berchtold et al. "Synaptic genes are extensively downregulated across multiple brain regions in normal human aging and Alzheimer's disease". In: *Neurobiology of Aging* 34.6 (June 2013), pp. 1653–1661. DOI: 10.1016/j.neurobiolaging.2012.11.024.

[70] Laura J. Blair et al. "Accelerated neurodegeneration through chaperone-mediated oligomerization of tau". In: *Journal of Clinical Investigation* 123.10 (Sept. 2013), pp. 4158–4169. DOI: 10.1172/jci69003.

[71] R. Edgar. "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". In: *Nucleic Acids Research* 30.1 (Jan. 2002), pp. 207–210. DOI: 10.1093/nar/30.1.207.

[72] Rafael A. Wu Zhijin; Irizarry. "A Model Based Background Adjustment for Oligonucleotide Expression Arrays". In: *Technical report, Johns Hopkins University, Dept. of Biostatistics Working Papers Working paper 1* (May 2004). URL: http://biostats.bepress.com/jhubiostat/paper1.

[73] Renesh Bedre. *reneshbedre/bioinfokit: Bioinformatics data analysis and visualization toolkit.* 2020. DOI: 10.5281/ZENODO.3965241.

[74] A. Voigt. *CSD. GitHub.* 2017. URL: https://github.com/andre-voigt/CSD.

[75] AlmaasLab Wiki. *HUNT Cloud — AlmaasLab Wiki.* [Online; accessed 25-Nov-2020]. 2019. URL: https://almaaslab.nt.ntnu.no/mediawiki/index.php?title=HUNT_Cloud&oldid=508.

[76] P. Shannon. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks". In: *Genome Research* 13.11 (Nov. 2003), pp. 2498–2504. DOI: 10.1101/gr.1239303.

[77] John D. Hunter. "Matplotlib: A 2D Graphics Environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/mcse.2007.55.

[78] M. Rebhan et al. "GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support". In: *Bioinformatics* 14.8 (Sept. 1998), pp. 656–664. DOI: 10.1093/bioinformatics/14.8.656.

[79] Ashburner et al. "Gene Ontology: tool for the unification of biology". In: *Nature Genetics* 25.1 (May 2000), pp. 25–29. DOI: 10.1038/75556.

[80] "The Gene Ontology Resource: 20 years and still GOing strong". In: *Nucleic Acids Research* 47.D1 (Nov. 2018), pp. D330–D338. DOI: 10.1093/nar/gky1055.

[81] Huaiyu Mi et al. "Large-scale gene function analysis with the PANTHER classification system". In: *Nature Protocols* 8.8 (July 2013), pp. 1551–1566. DOI: 10.1038/nprot.2013.092.

[82] Paul D. Thomas. "The Gene Ontology and the Meaning of Biological Function". In: *Methods in Molecular Biology*. Springer New York, Nov. 2016, pp. 15–24. DOI: 10.1007/978-1-4939-3743-1_2.

[83] Maxim V. Kuleshov et al. "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update". In: *Nucleic Acids Research* 44.W1 (May 2016), W90–W97. DOI: 10.1093/nar/gkw377.

[84] Katja Luck et al. "A reference map of the human binary protein interactome". In: *Nature* 580.7803 (Apr. 2020), pp. 402–408. DOI: 10.1038/s41586-020-2188-x.

[85] N Thirumoorthy. "Metallothionein: An overview". In: *World Journal of Gastroenterology* 13.7 (2007), p. 993. DOI: 10.3748/wjg.v13.i7.993.

[86] Philip L De Jager et al. "Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci". In: *Nature Neuroscience* 17.9 (Aug. 2014), pp. 1156–1163. DOI: 10.1038/nn.3786.

[87] Katie Lunnon et al. "Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease". In: *Nature Neuroscience* 17.9 (Aug. 2014), pp. 1164–1170. DOI: 10.1038/nn.3782.

[88] Wilhelm Stoffel and Andreas Bosio. "Myelin glycolipids and their functions". In: *Current Opinion in Neurobiology* 7.5 (Oct. 1997), pp. 654–661. DOI: 10.1016/s0959-4388(97)80085-2.

[89] Justin D. Lathia, Mark P. Mattson, and Aiwu Cheng. "Notch: from neural development to neurological disorders". In: *Journal of Neurochemistry* 107.6 (Dec. 2008), pp. 1471–1481. DOI: 10.1111/j.1471-4159.2008.05715.x.

[90] Guofeng Meng and Hongkang Mei. "Transcriptional Dysregulation Study Reveals a Core Network Involving the Progression of Alzheimer's Disease". In: *Frontiers in Aging Neuroscience* 11 (May 2019). DOI: 10.3389/fnagi.2019.00101.

[91] Lenz Steimer and Dagmar Klostermeier. "RNA helicases in infection and disease". In: *RNA Biology* 9.6 (June 2012), pp. 751–771. DOI: 10.4161/rna.20090.

[92] Karen Anthony and Jean-Marc Gallo. "Aberrant RNA processing events in neurological disorders". In: *Brain Research* 1338 (June 2010), pp. 67–77. DOI: 10.1016/j.brainres.2010.03.008.

[93] Tingting Yan, Feng Ding, and Yan Zhao. "Integrated identification of key genes and pathways in Alzheimer's disease via comprehensive bioinformatical analyses". In: *Hereditas* 156.1 (July 2019). DOI: 10.1186/s41065-019-0101-0.

[94] Ruijuan Xu et al. “Alkaline ceramidase 2 and its bioactive product sphingosine are novel regulators of the DNA damage response”. In: *Oncotarget* 7.14 (Mar. 2016), pp. 18440–18457. DOI: 10.18632/oncotarget.7825.

[95] Teja W. Groemer et al. “Amyloid Precursor Protein Is Trafficked and Secreted via Synaptic Vesicles”. In: *PLoS ONE* 6.4 (Apr. 2011). Ed. by Hitoshi Okazawa, e18754. DOI: 10.1371/journal.pone.0018754.

[96] W. Yang and S. Desiderio. “BAP-135, a target for Bruton’s tyrosine kinase in response to B cell receptor engagement”. In: *Proceedings of the National Academy of Sciences* 94.2 (Jan. 1997), pp. 604–609. DOI: 10.1073/pnas.94.2.604.

[97] G. Caraveo et al. “Action of TFII-I Outside the Nucleus as an Inhibitor of Agonist-Induced Calcium Entry”. In: *Science* 314.5796 (Oct. 2006), pp. 122–125. DOI: 10.1126/science.1127815.

[98] Amy N. Hicks et al. “Nicotinamide Mononucleotide Adenylyltransferase 2 (Nmnat2) Regulates Axon Integrity in the Mouse Embryo”. In: *PLoS ONE* 7.10 (Oct. 2012). Ed. by Stefan Strack, e47869. DOI: 10.1371/journal.pone.0047869.

[99] Tae Hyuk Kang et al. “The Housekeeping Gene Hypoxanthine Guanine Phosphoribosyltransferase (HPRT) Regulates Multiple Developmental and Metabolic Pathways of Murine Embryonic Stem Cell Neuronal Differentiation”. In: *PLoS ONE* 8.10 (Oct. 2013). Ed. by Austin John Cooney, e74967. DOI: 10.1371/journal.pone.0074967.

[100] Rachel Bell et al. “Systematic screening of the 14-3-3 eta chain gene for polymorphic variants and case-control analysis in schizophrenia”. In: *American Journal of Medical Genetics* 96.6 (2000), pp. 736–743. DOI: 10.1002/1096-8628-(20001204)96:6<736::aid-ajmg8>3.0.co;2-2.

[101] Felipe A Cisternas et al. “Cloning and characterization of human CADPS and CADPS2, new members of the Ca2-dependent activator for secretion protein family”. In: *Genomics* 81.3 (Mar. 2003), pp. 279–291. DOI: 10.1016/s0888-7543(02)00040-x.

[102] K. Nakanaga, K. Hama, and J. Aoki. “Autotaxin–an LPA producing enzyme with diverse functions”. In: *Journal of Biochemistry* 148.1 (May 2010), pp. 13–24. DOI: 10.1093/jb/mvq052.

[103] Shahzad Ahmad et al. “Association of lysophosphatidic acids with cerebrospinal fluid biomarkers and progression to Alzheimer’s disease”. In: *Alzheimer’s Research & Therapy* 12.1 (Oct. 2020). DOI: 10.1186/s13195-020-00680-9.

[104] Xin Yi, Xue-Jun Jiang, and Ze-Min Fang. “Histone methyltransferase SMYD2: ubiquitous regulator of disease”. In: *Clinical Epigenetics* 11.1 (Aug. 2019). DOI: 10.1186/s13148-019-0711-4.

[105] Nicolas Morel and Sandrine Poëa-Guyon. “The membrane domain of vacuolar HATPase: a crucial player in neurotransmitter exocytotic release”. In: *Cellular and Molecular Life Sciences* 72.13 (Mar. 2015), pp. 2561–2573. DOI: 10.1007/s00018-015-1886-2.

[106] Anna Kremer. “GSK3 and Alzheimer’s disease: facts and fiction...” In: *Frontiers in Molecular Neuroscience* 4 (2011). DOI: 10.3389/fnmol.2011.00017.

[107] Joana S. Cristóvão et al. “The S100B Alarmin Is a Dual-Function Chaperone Suppressing Amyloid- Oligomerization through Combined Zinc Chelation and Inhibition of Protein Aggregation”. In: *ACS Chemical Neuroscience* 11.17 (July 2020), pp. 2753–2760. DOI: 10.1021/acschemneuro.0c00392.

[108] Hana Antonicka et al. “The Mitochondrial RNA-Binding Protein GRSF1 Localizes to RNA Granules and Is Required for Posttranscriptional Mitochondrial Gene Expression”. In: *Cell Metabolism* 17.3 (Mar. 2013), pp. 386–398. DOI: 10.1016/j.cmet.2013.02.006.

[109] Theresa R. Bomfim et al. “An anti-diabetes agent protects the mouse brain from defective insulin signaling caused by Alzheimer’s disease–associated A oligomers”. In: *Journal of Clinical Investigation* 122.4 (Apr. 2012), pp. 1339–1353. DOI: 10.1172/jci57256.

[110] Konrad Talbot et al. “Demonstrated brain insulin resistance in Alzheimer’s disease patients is associated with IGF-1 resistance, IRS-1 dysregulation, and cognitive decline”. In: *Journal of Clinical Investigation* 122.4 (Apr. 2012), pp. 1316–1338. DOI: 10.1172/jci59903.

[111] Guini Hong et al. “Separate enrichment analysis of pathways for up- and down-regulated genes”. In: *Journal of The Royal Society Interface* 11.92 (Mar. 2014), p. 20130950. DOI: 10.1098/rsif.2013.0950.

[112] Ditte S. Jonesco, Morten A. Karsdal, and Kim Henriksen. “The CNS-specific proteoglycan, brevican, and its ADAMTS4-cleaved fragment show differential serological levels in Alzheimer’s disease, other types of dementia and non-demented controls: A cross-sectional study”. In: *PLOS ONE* 15.6 (June 2020). Ed. by Yona Levites, e0234632. DOI: 10.1371/journal.pone.0234632.

[113] Bhanu Jena. “Membrane Fusion: Role of SNAREs and Calcium”. In: *Protein & Peptide Letters* 16.7 (July 2009), pp. 712–717. DOI: 10.2174/0929866097886–81869.

[114] J. K. Choi et al. “Differential coexpression analysis using microarray data and its application to human cancer”. In: *Bioinformatics* 21.24 (Oct. 2005), pp. 4348–4355. DOI: 10.1093/bioinformatics/bti722.

[115] Antonio Reverter et al. “Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer”. In: *Bioinformatics* 22.19 (July 2006), pp. 2396–2404. DOI: 10.1093/bioinformatics/btl392.

[116] Marjan Farahbod and Paul Pavlidis. “Differential coexpression in human tissues and the confounding effect of mean expression levels”. In: *Bioinformatics* (July 2018). Ed. by Jonathan Wren. DOI: 10.1093/bioinformatics/bty538.

[117] Maite Solas, Elena Puerta, and Maria Ramirez. “Treatment Options in Alzheimer´s Disease: The GABA Story”. In: *Current Pharmaceutical Design* 21.34 (Oct. 2015), pp. 4960–4971. DOI: 10.2174/1381612821666150914121149.

[118] Karl H. Braunewell. “The visinin-like proteins VILIP-1 and VILIP-3 in Alzheimer’s disease—old wine in new bottles”. In: *Frontiers in Molecular Neuroscience* 5 (2012). DOI: 10.3389/fnmol.2012.00020.

[119] B. A. Hemmings and D. F. Restuccia. “PI3K-PKB/Akt Pathway”. In: *Cold Spring Harbor Perspectives in Biology* 4.9 (Sept. 2012), a011189–a011189. DOI: 10.1101/cshperspect.a011189.

[120] Amy R. Nelson et al. “Neurovascular dysfunction and neurodegeneration in dementia and Alzheimer’s disease”. In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1862.5 (May 2016), pp. 887–900. DOI: 10.1016/j.bbadis.2015.12.016.

[121] Qian Li and Miao Sun. “The role of autophagy in Alzheimer’s disease”. In: *Journal of Systems and Integrative Neuroscience* 3.4 (2017). DOI: 10.15761/jsin.1000172.

[122] George Bartzokis. “Alzheimer’s disease as homeostatic responses to age-related myelin breakdown”. In: *Neurobiology of Aging* 32.8 (Aug. 2011), pp. 1341–1371. DOI: 10.1016/j.neurobiolaging.2009.08.007.

[123] L. Liu-Yesucevitz et al. “Local RNA Translation at the Synapse and in Disease”. In: *Journal of Neuroscience* 31.45 (Nov. 2011), pp. 16086–16093. DOI: 10.1523/jneurosci.4105-11.2011.

[124] T. Matsuzaki et al. “Insulin resistance is associated with the pathology of Alzheimer disease: The Hisayama Study”. In: *Neurology* 75.9 (Aug. 2010), pp. 764–770. DOI: 10.1212/wnl.0b013e3181eee25f.

[125] Stefan F. Lichtenthaler. “Alpha-Secretase Cleavage of the APP: Proteolysis Regulated by Signaling Pathways and Protein Trafficking”. In: *Current Alzheimer Research* 9.2 (Feb. 2012), pp. 165–177. DOI: 10.2174/156720512799361655.

[126] Jo V. Rushworth and Nigel M. Hooper. “Lipid Rafts: Linking Alzheimer’s Amyloid-Production, Aggregation, and Toxicity at Neuronal Membranes”. In: *International Journal of Alzheimer’s Disease* 2011 (2011), pp. 1–14. DOI: 10.4061/2011/603052.

[127] Marine Mansuy et al. “Deletion of plasma Phospholipid Transfer Protein (PLTP) increases microglial phagocytosis and reduces cerebral amyloid- deposition in the J20 mouse model of Alzheimer’s disease”. In: *Oncotarget* 9.28 (Apr. 2018), pp. 19688–19703. DOI: 10.18632/oncotarget.24802.

[128] Mario Nizzari et al. “Neurodegeneration in Alzheimer Disease: Role of Amyloid Precursor Protein and Presenilin 1 Intracellular Signaling”. In: *Journal of Toxicology* 2012 (2012), pp. 1–13. DOI: 10.1155/2012/187297.

[129] Ramon Xulvi-Brunet and Hongzhe Li. “Co-expression networks: graph properties and topological comparisons”. In: *Bioinformatics* 26.2 (Nov. 2009), pp. 205–214. DOI: 10.1093/bioinformatics/btp632.

[130] Patrick Cahan et al. “CellNet: Network Biology Applied to Stem Cell Engineering”. In: *Cell* 158.4 (Aug. 2014), pp. 903–915. DOI: 10.1016/j.cell.2014.07.020.

[131] S. Ballouz, W. Verleyen, and J. Gillis. “Guidance for RNA-seq co-expression network construction and analysis: safety in numbers”. In: *Bioinformatics* 31.13 (Feb. 2015), pp. 2123–2130. DOI: 10.1093/bioinformatics/btv118.

[132] Nikolaos Koutsouleris et al. “Individualized differential diagnosis of schizophrenia and mood disorders using neuroanatomical biomarkers”. In: *Brain* 138.7 (May 2015), pp. 2059–2073. DOI: 10.1093/brain/awv111.

[133] Erika Kropf et al. “From anatomy to function: the role of the somatosensory cortex in emotional regulation”. In: *Brazilian Journal of Psychiatry* 41.3 (May 2019), pp. 261–269. DOI: 10.1590/1516-4446-2018-0183.

[134] Smyth G. et al. “Limma: linear models for microarray data”. In: *Bioinformatics and Computational Biology Solutions using R and Bioconductor* (2005), pp. 397–420.

[135] Janet Piñero et al. “The DisGeNET knowledge platform for disease genomics: 2019 update”. In: *Nucleic Acids Research* (Nov. 2019). DOI: 10.1093/nar/gkz1021.

Appendices

A Individual C-, S- and D-networks

The individual networks with exclusively one link type generated from the CSD framework were imported to Cytoscape and are visualized in Fig. 6.1-6.3. The network size of the C-, S- and D-networks are N = 331, N = 671 and N = 705, respectively.

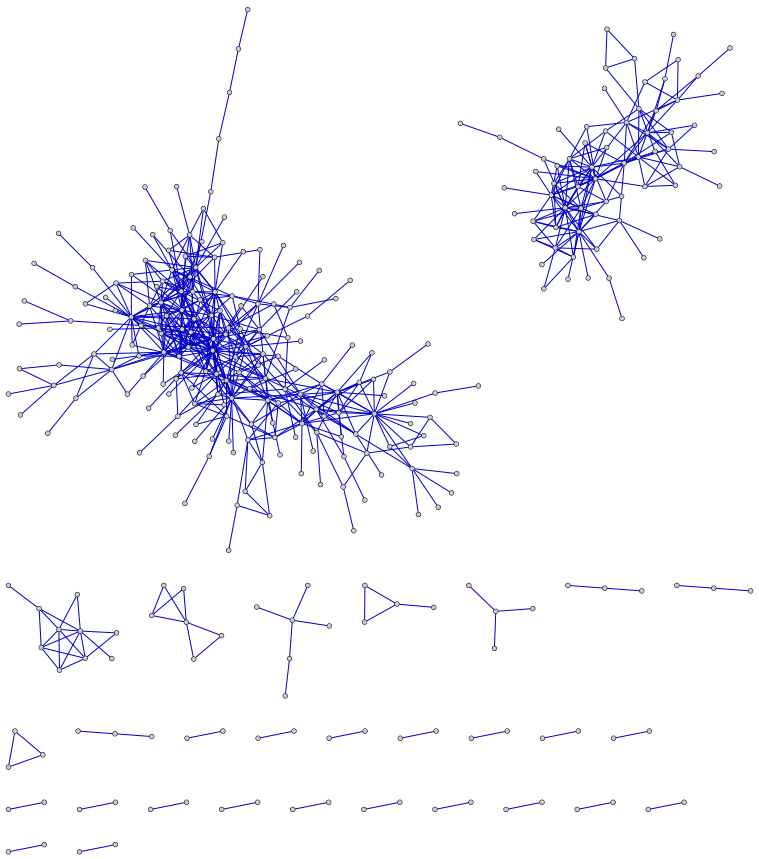


Figure 6.1: C-network, nodes represent genes and links their conserved type of co-expression. N = 331, M = 709.

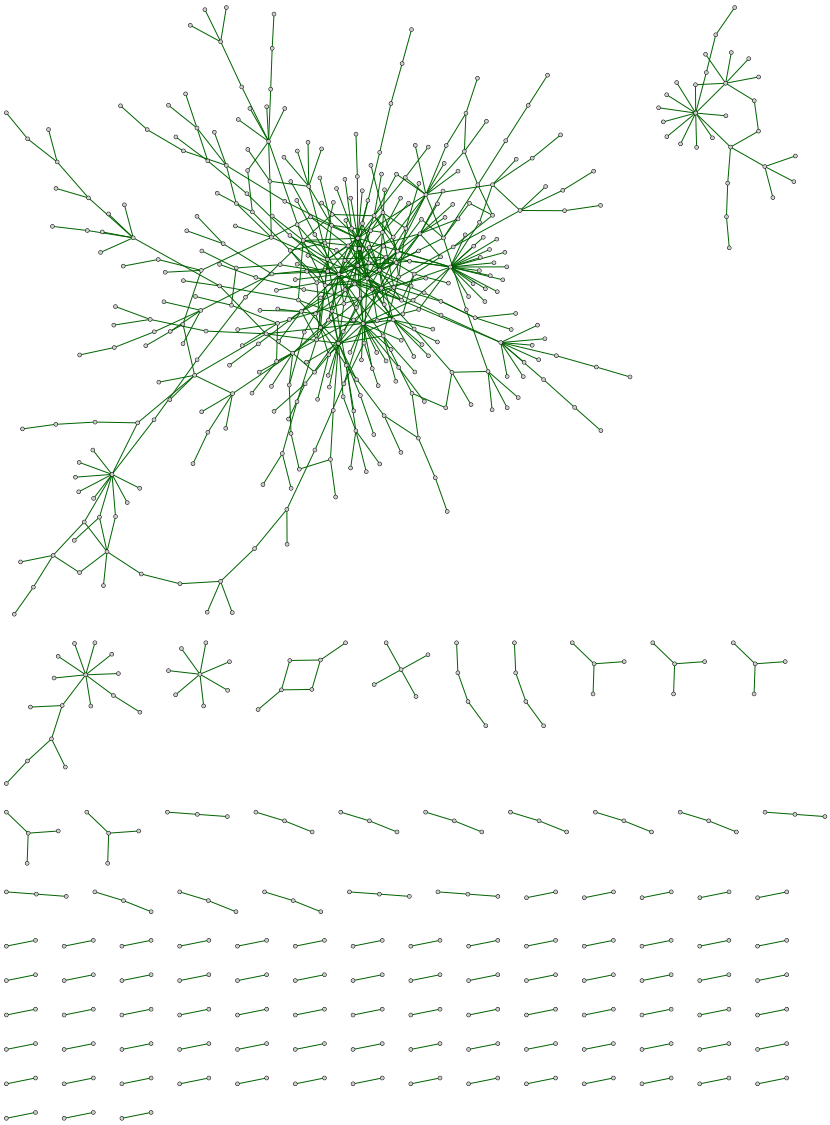


Figure 6.2: S-network, nodes represent genes and links their specific type of co-expression. N = 671, M = 690.

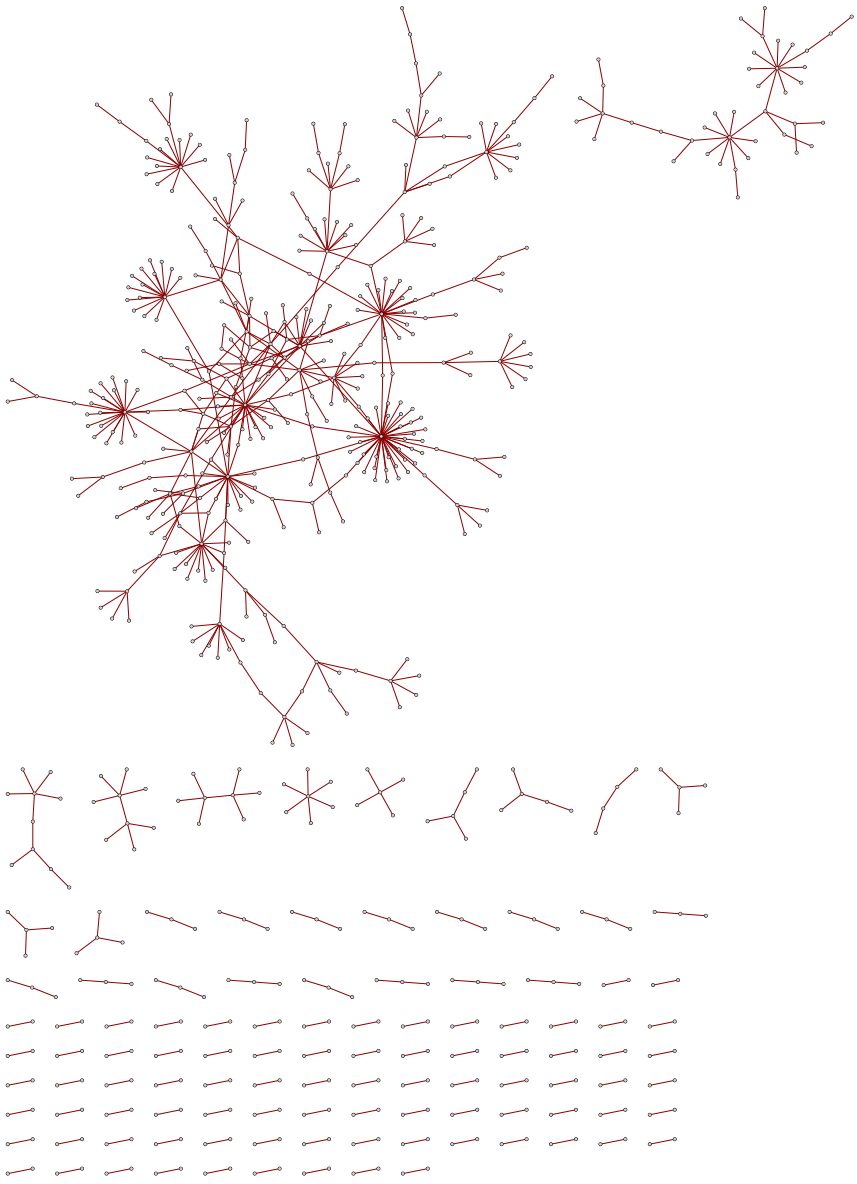


Figure 6.3: D-network, nodes represent genes and links their differentiated type of co-expression. N = 705, M = 645.

B Results from Enrichment Analyses

This appendix provides the complete results of the enrichment for GO biological processes and KEGG Pathways in different parts of the network that were too long for the Results section. All terms are considered significant with a BH-adjusted p-value (FDR) < 0.05, and sorted by Fold Enrichment (FE). Processes are over-represented compared to what could be expected to be drawn randomly from the database, unless stated otherwise. The number of genes enriched in the input list (#genes) and number of genes in reference list (#ref) is provided for all biological processes.

B.1 GO of S-network

Table 6.1 shows all significant terms enriched for the 671 genes with specific (S) co-expression. The result is sorted by hierarchy of the terms; by FE within each family of processes. FE above 1.0 represents an over-representation of genes, while values smaller than 1.0 represents an under-representation.

Table 6.1: All GO biological processes enriched in S-network. Sorted by fold enrichment (FE) within the hierarchy of the terms, most specific (child terms) first. #ref: number of genes in reference list. #genes: number of enriched genes in network. +/-: over/under-representation compared to expected. FDR: adjusted p-value by BH-method.

GO biological process	#ref	#genes	FE	+/-	FDR
plasma membrane bounded cell projection organization	1157	65	1.72	+	2.99E-02
cell projection organization	1203	67	1.71	+	3.00E-02
cellular component organization	5699	234	1.26	+	4.29E-02
cellular process	15632	573	1.12	+	1.34E-04
phosphate-containing compound metabolic process	2167	113	1.60	+	2.86E-03
phosphorus metabolic process	2194	114	1.59	+	3.47E-03
metabolic process	8585	338	1.21	+	1.44E-02
amide transport	1574	82	1.60	+	3.27E-02
transport	4572	204	1.37	+	3.14E-03
establishment of localization	4704	206	1.34	+	7.92E-03
localization	5862	249	1.30	+	4.10E-03
cellular protein localization	1646	84	1.56	+	4.37E-02
cellular localization	3007	146	1.49	+	2.95E-03
macromolecule localization	2564	126	1.51	+	7.34E-03
protein localization	2200	110	1.53	+	1.42E-02
regulation of localization	2784	139	1.53	+	1.94E-03
establishment of localization in cell	2378	116	1.50	+	1.72E-02
organic substance transport	2225	108	1.49	+	3.12E-02
regulation of signaling	3615	163	1.38	+	1.59E-02
regulation of cell communication	3576	160	1.37	+	2.48E-02
organonitrogen compound metabolic process	5450	230	1.29	+	1.48E-02
Unclassified	2858	46	.49	-	9.95E-05
detection of chemical stimulus involved in sensory perception of smell	439	1	.07	-	1.93E-02
detection of chemical stimulus involved in sensory perception	484	2	.13	-	2.80E-02
detection of stimulus involved in sensory perception	550	3	.17	-	2.69E-02
detection of chemical stimulus	520	2	.12	-	1.45E-02
sensory perception of smell	468	2	.13	-	4.34E-02

B.2 GO of C-modules

Tables 6.3-6.4 show the biological processes significantly enriched for each of the three modules with conserved co-expression. Only the most specific GO terms are shown for module 4 and 6, whereas for module 9 all significant terms are shown. The complete lists of terms for module 4 and 6 are provided in the following doi: 10.6084/m9.figshare.

Table 6.2: GO enrichment analysis of biological processes on module 6 (N = 92). Only the most specific terms are included, sorted by fold enrichment (FE). FDR: BH adjusted p-value.

GO biological process	#ref	#genes	FE	FDR
glutamate catabolic process to 2-oxoglutarate	2	2	>100	1.61E-02
glutamate catabolic process to aspartate	2	2	>100	1.60E-02
aspartate biosynthetic process	3	2	>100	2.37E-02
aspartate catabolic process	4	2	>100	3.35E-02
oxaloacetate metabolic process	8	3	87.86	2.44E-03
mitochondrial ATP synthesis coupled proton transport	22	4	42.60	1.05E-03
regulation of synaptic vesicle endocytosis	17	3	41.34	1.30E-02
cristae formation	32	5	36.61	2.75E-04
regulation of mitochondrial depolarization	21	3	33.47	1.98E-02
phagosome acidification	28	4	33.47	2.09E-03
transferrin transport	36	5	32.54	3.44E-04
gluconeogenesis	46	5	25.47	7.08E-04
organelle transport along microtubule	80	6	17.57	5.44E-04
synaptic vesicle cycle	117	6	12.01	2.82E-03
respiratory electron transport chain	110	5	10.65	1.90E-02
regulation of macroautophagy	177	7	9.27	2.73E-03
cellular response to insulin stimulus	177	6	7.94	1.91E-02
regulation of exocytosis	211	6	6.66	4.02E-02
vesicle localization	207	6	6.79	3.69E-02
regulation of neurotransmitter levels	216	6	6.51	4.46E-02
chemical synaptic transmission	414	9	5.09	1.30E-02
intracellular protein transport	992	13	3.07	3.89E-02
nervous system development	2203	21	2.23	4.50E-02

Table 6.3: GO enrichment analysis of biological processes on module 4 (N = 86). Only the most specific terms are included, sorted by fold enrichment (FE). FDR: Benjamini-Hochberg adjusted p-value.

GO biological process	#ref	#genes	FE	FDR
regulation of calcium ion-dependent exocytosis of neurotransmitter	3	2	>100	2.72E-02
regulation of synaptic activity	3	2	>100	2.70E-02
glutamine catabolic process	3	2	>100	2.67E-02
postsynaptic intermediate filament cytoskeleton organization	3	2	>100	2.65E-02
JUN phosphorylation	4	2	>100	3.52E-02
exocytic insertion of neurotransmitter receptor to postsynaptic membrane	4	2	>100	3.49E-02
synaptic vesicle membrane organization	4	2	>100	3.46E-02
regulation of synaptic vesicle priming	7	3	98.20	2.69E-03
glutamate biosynthetic process	5	2	91.65	4.38E-02
negative regulation of peptidyl-cysteine S-nitrosylation	5	2	91.65	4.35E-02
neurofilament cytoskeleton organization	8	3	85.92	3.43E-03
cardiac muscle hypertrophy in response to stress	14	3	49.10	1.08E-02
regulation of short-term neuronal synaptic plasticity	15	3	45.83	1.20E-02
positive regulation of calcium ion-dependent exocytosis	20	4	45.83	1.14E-03
calcium ion-regulated exocytosis of neurotransmitter	16	3	42.96	1.35E-02
glutamate secretion	31	5	36.96	1.96E-04
positive regulation of neurotransmitter secretion	19	3	36.18	1.90E-02
synaptic vesicle endocytosis	49	7	32.73	3.98E-06
regulation of vesicle fusion	23	3	29.89	2.67E-02
regulation of long-term neuronal synaptic plasticity	26	3	26.44	3.45E-02
regulation of amino acid transport	37	4	24.77	7.18E-03
dendritic spine development	29	3	23.70	4.19E-02
clathrin-dependent endocytosis	30	3	22.91	4.34E-02
negative regulation of G protein-coupled receptor signaling pathway	48	4	19.09	1.39E-02
cellular response to nerve growth factor stimulus	53	4	17.29	1.85E-02
associative learning	82	6	16.77	8.76E-04
vesicle docking	63	4	14.55	2.84E-02
vesicle fusion	80	5	14.32	8.19E-03
cellular response to calcium ion	85	5	13.48	1.03E-02
negative regulation of neuron apoptotic process	152	6	9.04	1.24E-02
positive regulation of neuron projection development	155	6	8.87	1.36E-02
negative regulation of neuron projection development	133	5	8.61	4.07E-02
import into cell	179	6	7.68	2.35E-02
regulation of endocytosis	205	6	6.71	3.86E-02
negative regulation of transport	449	9	4.59	2.52E-02
neuron projection morphogenesis	502	10	4.56	1.39E-02
cell morphogenesis involved in neuron differentiation	456	9	4.52	2.67E-02
head development	817	12	3.37	3.17E-02
cellular chemical homeostasis	749	11	3.37	4.80E-02
central nervous system development	1019	14	3.15	2.19E-02
response to organonitrogen compound	1001	13	2.98	4.62E-02

Table 6.4: GO enrichment analysis of biological processes on module 9 (N = 66). All significant terms are shown (Benjamini-Hochberg adjusted p-value (FDR) < 0.05), sorted by fold enrichment (FE).

GO biological process	#ref	#genes	FE	FDR
galactosylceramide biosynthetic process	6	3	>100	2.36E-03
glycosylceramide biosynthetic process	7	3	>100	2.99E-03
galactosylceramide metabolic process	9	3	>100	4.09E-03
galactolipid metabolic process	10	3	96.24	4.72E-03
central nervous system myelination	21	6	91.65	6.36E-07
glycosylceramide metabolic process	16	3	60.15	1.25E-02
oligodendrocyte development	46	6	41.84	2.57E-05
glycosphingolipid biosynthetic process	25	3	38.49	3.53E-02
oligodendrocyte differentiation	72	8	35.64	5.38E-07
ensheathment of neurons	112	10	28.64	6.81E-08
myelination	110	9	26.25	5.96E-07
ceramide biosynthetic process	51	4	25.16	1.24E-02
peripheral nervous system development	77	5	20.83	4.07E-03
glycosphingolipid metabolic process	63	4	20.37	2.48E-02
sphingolipid biosynthetic process	93	5	17.25	7.14E-03
glial cell development	115	6	16.74	2.35E-03
glial cell differentiation	180	9	16.04	1.56E-05
sphingolipid metabolic process	155	7	14.49	9.29E-04
regulation of gliogenesis	123	5	13.04	2.22E-02
gliogenesis	233	9	12.39	8.42E-05
membrane lipid biosynthetic process	133	5	12.06	3.00E-02
regulation of cell projection organization	710	10	4.52	2.97E-02
central nervous system development	1025	14	4.38	2.71E-03
neurogenesis	1703	21	3.96	4.75E-05
regulation of cellular component movement	1042	12	3.69	3.48E-02
regulation of hydrolase activity	1305	15	3.69	5.47E-03
generation of neurons	1599	16	3.21	1.26E-02
positive regulation of cellular protein metabolic process	1633	15	2.95	4.88E-02
cellular developmental process	3845	28	2.34	4.02E-03

B.3 GO of up-DEGs and down-DEGs

Table 6.5 and Table 6.6 show the most highly enriched biological processes found for the 497 up-regulated and 699 down-regulated genes in the AD transcriptomic data, respectively.

Table 6.5: GO biological processes enriched in up-DEGs. Only the most specific terms (w/ FDR < 0.05) are included, sorted by fold enrichment (FE). FDR: Benjamini-Hochberg adjusted p-value.

GO biological process	#ref	#genes	FE	FDR
peptide antigen assembly with MHC class II protein complex	4	4	42.90	1.79E-03
regulation of macrophage migration inhibitory factor signaling pathway	3	3	42.90	1.43E-02
regulation of type IIa hypersensitivity	4	3	32.18	2.15E-02
oligodendrocyte cell fate specification	5	3	25.74	3.02E-02
glomerular visceral epithelial cell migration	6	3	21.45	4.10E-02
regulation of germinal center formation	9	4	19.07	1.08E-02
positive regulation of intracellular estrogen receptor signaling pathway	12	4	14.30	2.17E-02
notochord development	20	5	10.73	1.41E-02
detection of external biotic stimulus	21	5	10.22	1.62E-02
cellular response to platelet-derived growth factor stimulus	21	5	10.22	1.62E-02
negative regulation of B cell activation	35	8	9.81	6.62E-04
amyloid-beta clearance	22	5	9.75	1.88E-02
negative regulation of Rho protein signal transduction	23	5	9.33	2.15E-02
negative regulation of interleukin-2 production	27	5	7.95	3.48E-02
regulation of lipopolysaccharide-mediated signaling pathway	27	5	7.95	3.47E-02
lung epithelium development	30	5	7.15	4.93E-02
astrocyte development	38	6	6.77	2.44E-02
cortical actin cytoskeleton organization	38	6	6.77	2.43E-02
glial cell activation	40	6	6.44	3.00E-02
glomerulus development	54	8	6.36	6.24E-03
regulation of interleukin-10 production	54	8	6.36	6.20E-03
positive regulation of B cell proliferation	42	6	6.13	3.50E-02
positive regulation of myeloid leukocyte differentiation	58	8	5.92	8.89E-03
intermediate filament-based process	52	7	5.78	2.12E-02
T cell differentiation in thymus	52	7	5.78	2.11E-02
neural precursor cell proliferation	75	10	5.72	2.25E-03
cell differentiation involved in kidney development	45	6	5.72	4.59E-02
positive regulation of tumor necrosis factor production	85	11	5.55	1.35E-03
positive regulation of phagocytosis	64	8	5.36	1.45E-02
positive regulation of response to cytokine stimulus	57	7	5.27	3.06E-02
regulation of osteoclast differentiation	67	8	5.12	1.80E-02

Table 6.6: GO biological processes enriched in down-DEGs. Only the most specific terms (w/ FDR < 0.05) are included, sorted by fold enrichment (FE). FDR: Benjamini-Hochberg adjusted p-value.

GO biological process	#ref	#genes	FE	FDR
maintenance of presynaptic active zone structure	4	4	29.45	6.44E-03
regulation of synaptic vesicle priming	7	5	21.04	2.55E-03
spontaneous neurotransmitter secretion	6	4	19.63	1.53E-02
regulation of short-term neuronal synaptic plasticity	15	8	15.71	8.83E-05
synaptic vesicle maturation	11	5	13.39	9.70E-03
calcium ion-regulated exocytosis of neurotransmitter	16	6	11.04	5.26E-03
positive regulation of calcium ion-dependent exocytosis	20	7	10.31	2.19E-03
neurotransmitter receptor internalization	15	5	9.82	2.57E-02
synaptic transmission, glutamatergic	31	10	9.50	1.23E-04
positive regulation of neurotransmitter secretion	19	6	9.30	1.01E-02
synaptic transmission, dopaminergic	16	5	9.20	3.10E-02
corpus callosum development	17	5	8.66	3.74E-02
phagosome acidification	28	8	8.41	1.96E-03
synaptic vesicle endocytosis	49	14	8.41	3.92E-06
transferrin transport	36	10	8.18	3.01E-04
positive regulation of excitatory postsynaptic potential	29	8	8.12	2.30E-03
regulation of AMPA receptor activity	27	7	7.64	8.24E-03
glutamate secretion	31	8	7.60	3.22E-03
regulation of synaptic vesicle recycling	24	6	7.36	2.49E-02
axon extension	36	8	6.54	7.01E-03
vesicle docking involved in exocytosis	42	9	6.31	3.66E-03
long-term synaptic potentiation	48	10	6.14	1.96E-03
response to morphine	35	7	6.14	2.61E-02
long-term memory	35	7	5.89	2.60E-02
regulation of dopamine secretion	36	7	5.73	2.92E-02
neuron recognition	49	9	5.41	8.71E-03
regulation of postsynaptic membrane neurotransmitter receptor levels	39	7	5.29	4.12E-02
receptor localization to synapse	39	7	5.29	4.11E-02
regulation of sodium ion transmembrane transporter activity	56	10	5.26	5.03E-03

B.4 KEGG Pathways in Module 6

The complete results from the KEGG Pathways enrichment of module 6, including all gene symbols for each term is found in the following doi: 10.6084/m9.figshare. Table 6.7 shows all the significantly enriched pathways after correcting for multiple testing (FDR < 0.05).

Table 6.7: All significantly enriched KEGG Pathways in module 6, sorted by Fold Enrichment (FE). FDR: Benjamini-Hochberg adjusted p-value.

Term	FE	FDR
Phenylalanine, tyrosine and tryptophan biosynthesis	93.0	5.07E-03
Collecting duct acid secretion	34.5	2.28E-04
Phenylalanine metabolism	27.4	0.0334
Arginine biosynthesis	22.2	0.0400
Synaptic vesicle cycle	20.9	2.70E-06
Epithelial cell signaling in Helicobacter pylori infection	20.5	2.40E-05
Oxidative phosphorylation	19.2	3.65E-09
Vibrio cholerae infection	18.6	2.16E-03
Parkinson disease	18.0	3.73E-09
Cysteine and methionine metabolism	14.8	0.0176
Alzheimer disease	13.6	2.50E-07
Huntington disease	13.3	6.79E-08
VEGF signaling pathway	11.8	0.0308
Long-term potentiation	10.4	0.0372
Fc gamma R-mediated phagocytosis	10.2	0.0123
Rheumatoid arthritis	10.2	0.0116
Renin secretion	10.1	0.0374
Renal cell carcinoma	10.1	0.0388
T cell receptor signaling pathway	9.21	0.0162
Cardiac muscle contraction	8.94	0.0460
cGMP-PKG signaling pathway	8.41	2.50E-03
Phagosome	7.65	0.0119
mTOR signaling pathway	7.65	0.0129
Oxytocin signaling pathway	7.60	0.0114
Oocyte meiosis	7.44	0.0320
Cellular senescence	7.27	0.0131
Thermogenesis	7.05	2.39E-03
Non-alcoholic fatty liver disease (NAFLD)	6.24	0.0416
Human T-cell leukemia virus 1 infection	5.31	0.0342
Endocytosis	4.77	0.0416
Human papillomavirus infection	4.23	0.0385

C Biological functions of network hubs

Table 6.8 shows the 18 network hubs identified with the biological function of their gene product, as annotated from GeneCards [78]. These all have 20 or more neighbors in the CSD network, indicating a prominent role for the network topology.

Table 6.8: Largest hubs in the CSD network and the associated biological function of their gene product (mostly proteins). Genes are colored according to the predominant link type (C = blue, S = green and D = red).

Gene	Function
KIAA1841	Uncharacterized protein
NMNAT2	Nicotinamide Mononucleotide Adenylyltransferase 2. Cytosolic enzyme that catalyzes the formation of NAD+ from nicotinamide mononucleotide (NMN) and ATP.
MIGA2	Mitoguardin 2. Regulator of mitochondrial fusion: acts by forming homo- and heterodimers at the mitochondrial outer membrane and facilitating the formation of PLD6/MitoPLD dimers. May act by regulating phospholipid metabolism via PLD6/MitoPLD. (PLD6 gene codes for Mitochondrial cardiolipin hydrolase)
AQR	RNA helicase. Aquarius Intron-Binding Spliceosomal Factor. Involved in pre-mRNA splicing as component of the spliceosome.
AL158206.1	Long non-coding RNA (lncRNA). Transcript overlaps with *ACER2* (gene coding for the protein Alkaline ceramidase 2, which catalyzes the hydrolysis of ceramide into sphingosine and free fatty acids at alkaline pH.
HPRT1	Hypoxanthine Phosphoribosyltransferase 1. Catalyzes the conversion of hypoxanthine to inosine monophosphate and guanine to guanosine monophosphate (nucleotide metabolism).
GTF2I	Gene that codes for two proteins: TFII-1 and BAP-135. TFII-1 is a general transcription factor which regulates gene activity by binding to promoter elements. BAP-135 is active in B-cells where it contributes to normal immune system function.
TOM1L2	Target Of Myb1 Like 2 Membrane Trafficking Protein. Belongs to a family of TOM1-related proteins involved in vesicular trafficking through the endocytic pathway. It recruits clathrin onto endosomes and modulates endosomal function.
YWHAH	Tyrosine 3-Monooxygenase/Tryptophan 5-Monooxygenase Activation Protein (eta isoform). Belongs to 14-3-3 family which mediate signal transduction by binding to and activating phosphoserine-containing proteins.
GOT1	Glutamic-Oxaloacetic Transaminase 1. Pyridoxal phoshate-dependent enzyme which synthesizes L-glutamate from L-aspartate or L-cysteine in the cytosol (amino acid metabolism).
NAPB	N-ethylmaleimide-sensitive factor (NSF) Attachment Protein Beta. Part of the 20S NSF-SNAP-SNARE complex, required for vesicular transport between the endoplasmic reticulum (ER) and Golgi apparatus.
TMEM178A	Transmembrane Protein 178A. Negative regulator of osteoclast differentiation in basal and inflammatory conditions by regulating Ca^{2+}-fluxes.
PLTP	Phospholipid Transfer Protein. Transfers phospholipids and free cholesterol from low density lipoproteins (LDL) and very low density lipoproteins (VLDL) into high-density lipoproteins (HDL).
LCAT	Lecithin-Cholesterol Acyltransferase. Glycoprotein which converts free cholesterol into cholesteryl esters on the surface of lipoproteins, resulting in mature spherical HDL.
ENPP2	Ectonucleotide Pyrophosphatase/Phosphodiesterase 2, also called Autotaxin. Functions both as a phosphodiesterase, which cleaves phosphodiester bonds at the 5' end of oligonucleotides, and a phospholipase, which catalyzes production of lysophosphatidic acid (LPA) in extracellular fluids. LPA evokes growth factor-like responses including stimulation of cell proliferation and chemotaxis.
CADPS	Calcium Dependent Secretion Activator. Neural/endocrine-specific membrane protein required for the Ca^{2+}-regulated exocytosis of secretory vesicles, which also involves the synthesis of phosphatidylinositol 4,5-bisphosphate (PtdIns(4,5)P2).
MDH1	Malate Dehydrogenase 1. Catalyzes the NAD/NADH-dependent, reversible oxidation of malate to oxaloacetate in many metabolic pathways, including the citric acid cycle. Cytosolic isozyme, which plays a key role in the malate-aspartate shuttle that allows malate to pass through the mitochondrial membrane to be transformed into oxaloacetate for further cellular processes.
VSNL1	Visinin Like 1. Member of visinin/recoverin subfamily of neuronal calcium sensor proteins. Modulates intracellular signaling pathways of the central nervous system by regulating the activity of adenylyl cyclase.

D Python scripts for DEA

The following code is the self-written Python script used for the identification of DEGs and their statistical significance in the microarray data.

Listing 6.1: This script calculates mean expression of all genes in control and case patients separately. Then calculates $\log_2$ fold change, raw p-value, T-statistic and adjusted p-value (FDR by Benjamini-Hochberg). DF: Dataframe (Pandas). User must change names of input and output files, corresponding to desired sample set calculation.

```
import pandas as pd
import numpy as np

##Read expression data for control samples to DF##
data2 = pd.read_csv('ADcontrolold.txt', sep = '\t', header = None, skiprows=[0], index_col=0)
df2 = pd.DataFrame(data2)
df2['mean'] = df2.mean(axis=1) #mean expression of each row (gene)

##Expression for AD patients##
data3 = pd.read_csv('ADexp.txt', sep = '\t', header = None, skiprows=[0], index_col=0)
df3 = pd.DataFrame(data3)
df3['mean'] = df3.mean(axis=1)

##Calculate fold change from control to AD##
log2FC = (df3['mean']-df2['mean']) #FC is difference in mean, values already log2
newdf = pd.DataFrame(index=df2.index.copy()) #new DF with gene name as index column
newdf.index.names = ['Gene_name'] #Gene name as column name
newdf['Mean_control_exp.'] = round(df2['mean'],2) #control mean exp. to new column, 2 decimals
newdf['Mean_AD_exp.'] = round(df3['mean'],2) #add mean expression for sick as new column
newdf['log2FC'] = round(log2FC,2) #add log2FC as new column to DF

##Perform multiple t-tests, row-by-row (for every gene in both DFs)##
from scipy.stats import ttest_ind #independent t-test
df_m = pd.merge(df2, df3, left_index=True, right_index=True)
T_stat, p_vals = ttest_ind(df_m.iloc[:, df2.shape[1]:-1], df_m.iloc[:, :df2.shape[1]-1], axis=1)
newdf['Raw_p-value'] = np.round(p_vals, decimals = 4) #add p-values w/ 4 decimals to DF
newdf['T-statistic'] = np.round(T_stat, decimals = 2) #add T-statistic w/ 2 decimals to DF

#Function that adjusts p-vals, returns Benjamini-Hochberg adjusted P-value
def fdr(p_vals):
    from scipy.stats import rankdata
    ranked_p_values = rankdata(p_vals)
    fdr = p_vals * len(p_vals) / ranked_p_values
    fdr[fdr > 1] = 1
    return fdr

newdf['FDR'] = np.round(fdr(p_vals), decimals = 2) #perform function and add adjusted P-value

#Write DF to file, sorted from highest to lowest fold change
newdf.sort_values(by=['log2FC'], ascending = False, inplace = True)
newdf.to_csv('diffstats_allregionsold.txt', index=True, sep = '\t')
```

Listing 6.2: This script opens the file generated in the previous listing, with the results from the differential expression analysis. It prints the number of up-DEGs and down-DEGs and adds the gene symbols of these to individual files for easy incorporation into functional annotation.

```
#Opens file with results from the differential expression analysis
with open('diffstats_allregionsold.txt') as f:
    upDEGs = []
    downDEGs = []
    firstline = f.readline()
    for line in f:
        splitline = line.rstrip().split('\t')
        if float(splitline[6])<0.05: #genes with FDR < 0.05 are significant
            if float(splitline[3])>=0.2: #log2FC > 0.2 are upregulated
                upDEGs.append(splitline[0]) #add gene symbol
            elif float(splitline[3])<=(-0.2): #log2FC < -0.2 are downregulated
                downDEGs.append(splitline[0]) #add gene symbol
print(len(upDEGs), upDEGs) #Number of up-regulated genes and the list of gene symbols
print(len(downDEGs), downDEGs) #Number of down-regulated genes and the list of gene symbols

#Create a file where the gene symbol of up-DEGs are on invidual lines
with open('up-DEGs.txt', 'w') as file1:
    for i, gene in enumerate(upDEGs):
        file1.write(gene)
        file1.write('\n')

#Create a file where the gene symbol of down-DEGs are on invidual lines
with open('down-DEGs.txt', 'w') as file2:
    for i, gene in enumerate(downDEGs):
        file2.write(gene)
        file2.write('\n')
```

E List of DEGs in CSD network

This appendix includes the full list of the 229 genes recognized in the CSD network as differentially expressed genes (DEGs), split into up-DEGs and down-DEGs (Table 6.9).

Table 6.9: All DEGs recognized in the CSD network. 104 up-regulated genes and 125 down-regulated genes. Genes are listed from the largest to the smallest magnitude of change (absolute log2FC). Genes previously associated with AD are marked in bold.

	Gene symbol
Up-regulated genes	ADAMTS2 MYBPC1 ANGPT1 VAC14-AS1 HLA-DRB1 ITPKB GMPR NFKBIA CD74 SELL ITGB8 HCLS1 DOCK5 BBOX1 S1PR3 HSPA2 DNALI1 NACC2 PRKX SLC38A2 AC005332.4 C5AR1 H1-2 RFX4 H2BC19P CSRP1 TP53INP1 CHST6 C1orf87 ZFP36L2 CGNL1 ZCCHC24 MAP4K4 **GFAP** **TLR2** CHD7 POU3F2 MTMR10 RIN2 ST6GALNAC3 PANTR1 RNF130 KIAA1958 EZR LIFR RDX CCDC69 ANP32B PLEKHA7 EMX2OS DIAPH3 CTSH ANO6 AFF1 CRB1 ADGRA3 PTPN21 MMP8 ZNF423 C21orf62 ARHGAP42 USP54 CRB2 HIPK2 MYO10 LRP4 GLIS3 TEX26 STEAP3 GOLIM4 CDC42EP4 **KLC1** PLXNB1 TOB1 SERPINI2 HMG20B TMEM47 NDE1 CCDC151 IKBKB HEATR5A IQCK RNF19A FRYL PRKG1 NXT2 OR7E14P RFX2 USP53 CXCL16 WDR49 CERS1 PLXNB2 PELI2 ANAPC16 TAB2 NOTCH2 TNS2 GPRC5B HDAC1 FBXL7 PTBP1 SAP30L VAMP3
Down-regulated genes	RGS4 MAL2 OLFM3 KCNV1 RAB3C CDC42 NAP1L5 GABRG2 RTN4RL2 GAD1 SYN2 NMNAT2 **STMN2** CADPS SYT13 C3orf80 BRWD1 GNG3 ATP6V1G2 SNX10 SEZ6L2 PCLO BEX5 PAK1 YWHAH NSF TCERG1L HSPB3 **SNCB** CPNE4 HPRT1 SYNGR3 **SYP** AMPH RNF128 KALRN ACOT7 **UCHL1** SYN1 DYNC1I1 ATP6V1B2 ATP8A2 EEF1A2 NECAP1 AP3B2 SCN2A JPT1 LYRM9 TMEM178B UBE2T CLSTN2 GOT1 **CALM3** SV2B SYT1 MLLT11 BEX1 DNM1 CLSTN3 **PLD3** RFPL1S PGAP4 SYNGR1 STX1B PHF24 NAPB TAGLN3 RASAL1 GPRASP1 MOAP1 ENO2 GPRASP2 KLF10 SCG5 ADAM23 TBC1D9 ATP6V1A LYNX1 ITFG1 STXBP1 NDRG4 SLITRK3 MAPRE3 EID2 RUNDC1 DNAJC5 RNF41 FBLL1 SGIP1 AC139256.2 AFF2 INPP5F SV2A GPI SCAMP5 BCAT1 SULT4A1 **VDAC1** AL031118.1 **HMGCR** CD200 KIF3A AP2M1 MRPL15 GLT1D1 ATP5MC3 ATP6V1E1 AC005229.4 TFRC NKD2 YWHAZ BTBD10 AC006058.1 ATP5F1B PEX3 TOMM20 RAN PEX11B RER1 ADAM11 TMEM178A APOO WDR47 CCDC32 WDR74

A large number of the downregulated genes in Table 6.9 were found in module 4 and module 6:

- Module 4: RGS4, **STMN2**, SYT13, GNG3, **SNCB**, SYNGR3, **SYP**, KALRN, SYN1, ATP8A2, EEF1A2, SCN2A, SV2B, SYT1, DNM1, SYNGR1, STX1B, PHF24, NAPB, TAGLN3, GPRASP1, ENO2, GPRASP2, STXBP1, FBLL1, INPP5F, SV2A, SCAMP5, SULT4A1, CD200 and GLT1D1.
- Module 6: MAL2, OLFM3, KCNV1, RAB3C, CDC42, NAP1L5, GABRG2, SYN2, NMNAT2, CADPS, C3orf80, ATP6V1G2, SEZ6L2, BEX5, PAK1, YWHAH, NSF, HPRT1, AMPH, ACOT7, **UCHL1**, DYNC1I1, ATP6V1B2, NECAP1, AP3B2, TMEM178B, GOT1, **CALM3**, MLLT11, BEX1, **PLD3**, PGAP4, MOAP1, SCG5, ADAM23, TBC1D9, ATP6V1A, ITFG1, NDRG4, MAPRE3, EID2, DNAJC5, RNF41, GPI, **VDAC1**, AP2M1, MRPL15, ATP5MC3, ATP6V1E1, BTBD10, ATP5F1B, TOMM20, RAN and PEX11B.

F CSD network from complete data set

The complete data set from E-GEOD-48350 [67] included 80 patients with AD (60-95 years) and 173 healthy controls (20-99 years). As seen, the age span for controls was substantially higher than for case. Age was considered the most important potential confounding factor in the microarray data set. Fig. 6.4 shows the network resulting from implementing the CSD method on the full data set. Initially, we see that the network is more dense and that the separate components seen in the main CSD network (Fig. 4.1) are here all connected in the giant component. We can recognize some of the same hubs, but there are also some hubs with substantially higher degrees (see table 6.10). Together, the network hubs and their nearest neighbors represent as much as 45.4 % of all the links in the network. It would be interesting to study the roles of the largest hubs and compare the two CSD networks in order to indicate age-related gene expression. This was however considered outside the scope of this thesis, and is left for future work.

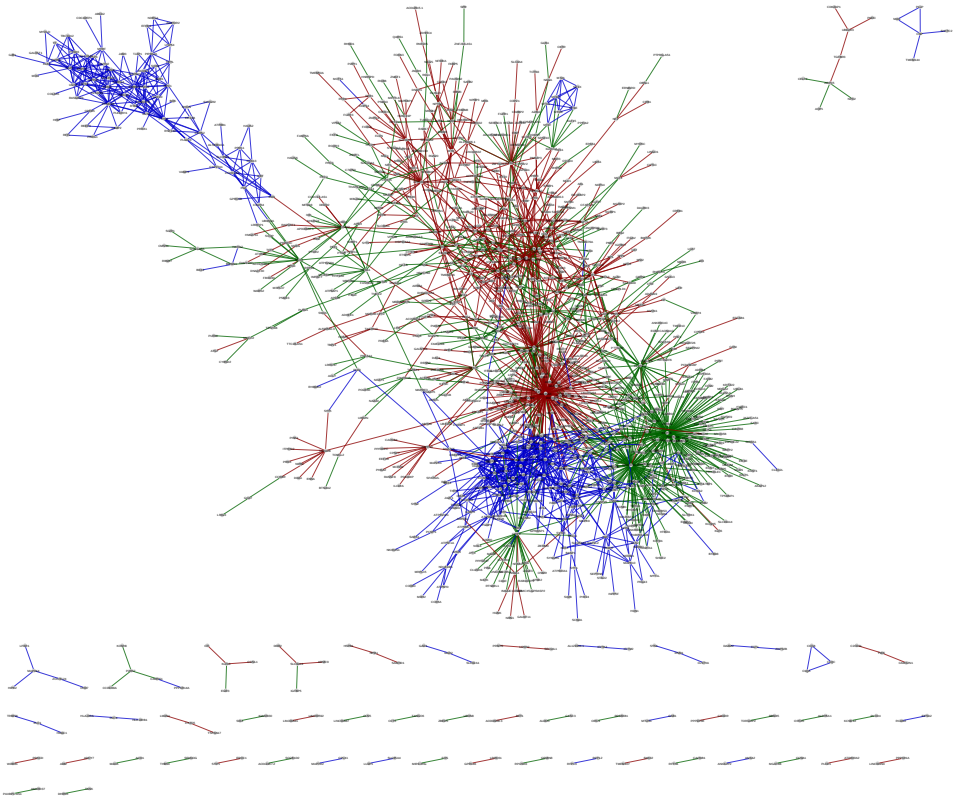


Figure 6.4: CSD network from full microarray data (80 AD patients and 173 controls). $N = 1230$, $M = 2072$. Nodes represent genes and links represent the type of co-expression between pairs of genes. Links are colored by type: blue is conserved (C), green is specific (S) and red is differentiated (D). Network generated using an importance level of $p = 5 \cdot 10^{-6}$ and visualized in *Cytoscape*.

Table 6.10: Network hubs and their degree in the CSD network from complete microarray data (AD = 80, Control = 173). Node degree k>=20 identified as hubs. Gene symbols are colored based on the predominant link type it has to its nearest neighbors: blue = C, green = S, red = D.

Gene	***k***
SLC46A3	166
EXPH5	145
AP000766.1	96
GRSF1	63
HPRT1	41
CYS1	39
NF1	34
CLASP1	34
CRIP2	32
NMNAT2	31
LRIG1	29
AC009407.1	26
RTN1	25
CADPS	24
LINC00173	24
GOT1	23
GTF2I	23
KIFAP3	22
ERBB3	21
ATP6V1B2	20
ATP6V1G2	20
MDH1	20
MIGA2	20

G DEA on specific brain tissue regions

DEA was performed on each of the 4 brain regions within the AD transcriptomic data to look for larger expression changes than what was seen for the pooled samples used in the main analysis. The number of samples (AD/control) for hippocampus (HC), entorhinal cortex (EC), superior frontal gyrus (SFG) and postcentral gyrus (PCG) were:

- HC: 19/25
- EC: 15/18
- SFG: 21/26
- PCG: 25/24

Significant differential expression after multiple testing correction was only found for the hippocampus region ($q = 0.04$). This strengthens the hypothesis that the transcriptional regulation in AD might be tissue-specific, at least for this data set. Table 6.11 shows that the top DEGs in HC have larger $\log_2$FC-values, especially the downregulated genes. Specifically, the most downregulated gene was *CALB1* had more than a 3-fold decrease in expression from control to case ($\log_2$FC = -1.81, $p = 0.0002$). This gene and *SST* were also top 5 down-DEGs for the overall brain tissue as well, but with lower magnitudes of change (Table 4.8).

Table 6.11: Top 5 up-DEGs and down-DEGs in hippocampus of individuals >= 60 years in Alzheimer's dataset, sorted by $\log_2$FC ($\log_2$ Fold Change). Mean gene expression is transformed with logarithm base 2 ($\log_2$). AD: Alzheimer's disease. FDR = Benjamini-Hochberg adjusted p-value.

Gene	Control exp.	AD exp.	$\log_2$FC	Raw p-value	T-statistic	FDR
CP	7.99	9.07	1.08	0.0004	3.88	0.04
CRLF1	7.57	8.44	0.88	0.0005	3.75	0.04
ANGPT1	7.54	8.37	0.84	0.0009	3.58	0.04
TNFRSF11B	5.14	5.91	0.77	0.0002	4.06	0.04
CAPS	7.92	8.67	0.75	0.0009	3.57	0.04
MAL2	8.56	7.19	-1.37	0.0003	-3.91	0.04
SST	8.73	7.16	-1.56	0.0001	-4.34	0.04
CHGB	9.32	7.73	-1.6	0.0001	-4.2	0.04
TAC1	7.69	5.9	-1.79	0.0002	-4.13	0.04
CALB1	7.84	6.03	-1.81	0.0002	-4.06	0.04

No statistically significant DEGs were found in EC or SFG. However, the genes with the highest magnitudes of change were LINC01094 (log2FC = 1.16, $p = 0.0113$, $q = 0.28$) and ADAMTS2 (log2FC = 1.05, $p = 0.0001$, $q = 0.09$), respectively. The brain region PCG seemed to be the least affected by the disease; virtually all "DEGs" were false positives ($q = 0.99$).

Susanne Glenna

Network analysis of Alzheimer's disease gene expression

NTNU
Norwegian University of
Science and Technology