Master's thesis

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

Anna Sunniva Hessevik

Differential Gene Co-expression Analysis of Bipolar Disorder Using a Consensus Network Approach

Master's thesis in Biotechnology (MBIOT5)
Supervisor: Eivind Almaas
Co-supervisor: André Voigt
May 2022

NTNU
Norwegian University of
Science and Technology

Anna Sunniva Hessevik

Differential Gene Co-expression Analysis of Bipolar Disorder Using a Consensus Network Approach

Master's thesis in Biotechnology (MBIOT5)
Supervisor: Eivind Almaas
Co-supervisor: André Voigt
May 2022

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

NTNU
Norwegian University of
Science and Technology

Summary

Differential gene co-expression network analysis has recently emerged as an important strategy for identification and investigation of potential dysregulated genes and pathways in diseases and other conditions. The Conserved, Specific, Differentiated (CSD) approach, a systematic framework for differential co-expression network analysis, defines three different types of co-expression between genes: conserved (C), specific (S) and differentiated (D). Here, the CSD approach will be utilized to investigate alterations in gene co-expression patterns in bipolar disorder (BP). As most BP data sets are characterized by small sample sizes, this thesis has also developed and tested methods for combining several data sets into one consensus CSD network. In general, this may be viewed as an extension of the current CSD approach as a method for dealing with small sample sizes.

Two approaches have been tested for creation of consensus CSD networks, which both combine Spearman rank correlation coefficients from individual data sets into averaged values that may be used as input to the conventional CSD approach. The first method relies on averages of Fisher's Z transformed correlation coefficients, while the second method utilizes weighted untransformed averages of correlation coefficients. The two methods generated comparable combined scores and gave rise to similar networks. Moreover, both combination methods outperformed the current approach for dealing with small sample sizes in CSD analysis. However, the method based on weighted untransformed averages appeared to be most suitable for CSD analysis as it is more conservative and less affected by spurious perfect correlations than Fisher's Z transformed averages. Structural analyses of a consensus CSD network also indicated that this method produces networks with many typical characteristics of conventional CSD networks.

The newly developed consensus CSD approach, based on weighted untransformed averages of correlation coefficients, was used to examine potential alterations in gene co-expression patterns in BP by combining six data sets. The functional analyses of the generated network suggested a potential role for mislocalization of proteins as at least three of the central genes in the network (SRP9, SRP14, GOLPH3L and possibly RBM23) are involved in this process and appeared to be dysregulated. In addition, the functional analyses suggested alterations in the specification process in the dorsolateral prefrontal cortex (DLPFC) of BP patients. This was supported by identification of at least two potentially dysregulated hubs with roles in this process (PITX3, FBLN2 and possibly RBM23), a gain of correlations for BP patients compared to control samples in the majority of S links as well as a suggested shift of the BP DLPFC correlations towards correlations from basal ganglia. It should be noted that the roles of mislocalization and alterations of specification processes were not supported by Gene Ontology (GO) analyses, illustrating that further investigation of BP is required.

Sammendrag

Analyser av differensielle gen-samuttrykksnettverk har nylig blitt utviklet som en viktig strategi for identifisering og undersøkelse av potensielt dysregulerte gener og spor i sykdommer og andre tilstander. CSD-metoden, et systematisk rammeverk for analyser av differensielle samuttrykksnettverk, definerer tre ulike typer samuttrykk mellom gener: konservert (C), spesifikk (S) og differensiert (D). Her benyttes CSD-metoden for å undersøke endringer i gen-samuttrykksmønstre i bipolar lidelse (BP). Ettersom mange BP datasett er karakterisert av en liten prøvestørrelse, har denne oppgaven også utviklet og testet metoder for å kombinere flere datasett til et konsensus CSD nettverk. Dette kan også ansees som en generell utvidelse av den nåværende CSD-metoden for å tilpasse fremgangsmåten til små prøvestørrelser.

To fremgangsmåter for å lage konsensus CSD nettverk har blitt testet i denne oppgaven. Begge metodene kombinerer Spearman rang korrelasjonskoeffisienter fra individuelle datasett til gjennomsnittlige verdier som kan benyttes som input til den konvensjonelle CSD-metoden. Den første fremgangsmåten er basert på gjennomsnitt av Fishers Z transformerte korrelasjonskoeffisienter, mens den andre metoden benytter vektede utransformerte gjennomsnitt av korrelasjonskoeffisienter. De to fremgangsmåtene ga sammenlignbare kombinerte verdier og liknende nettverk. I tillegg utkonkurrerte begge metodene den nåværende tilnærmingen for å håndtere små prøvestørrelse i CSD analyser. Metoden basert på vektede utransformerte gjennomsnitt virket imidlertid for å være mest passende for CSD analyse ettersom den er mer konservativ og mindre påvirket av tilfeldige perfekte korrelasjoner enn Fishers Z transformerte gjennomsnitt. Strukturelle analyser av et konsensus CSD nettverk indikerte også at denne metoden produserer nettverk med mange typiske egenskaper for konvensjonelle CSD nettverk.

Den nylig utviklede konsensus CSD-metoden, basert på vektede utransformerte gjennomsnitt av korrelasjonskoeffisienter, ble brukt for å undersøke potensielle endringer i gen-samutrykksmønstre i BP ved å kombinere seks datasett. Funksjonelle analyser av det genererte nettverket antydet en potensiell rolle for feillokalisering av proteiner ettersom minst tre av de sentrale genene i nettverket (SRP9, SRP14, GOLPH3L og muligens RBM23) er involverte i denne prosessen og så ut til å være dysregulerte. De funksjonelle analysene antydet i tillegg en endring i spesifiseringsprosessen i den dorsolaterale prefrontale cortex (DLPFC) hos BP pasienter. Dette var støttet av identifisering av minst to potentielt dysregulerte nettverksnav med roller i denne prosessen (PITX3, FBLN2 og muligens RBM23), en tilegning av korrelasjoner for BP pasienter sammenliknet med kontroll prøver for majoriteten av S linker og en mulig forskyving av korrelasjoner i DLPFC hos BP pasienter mot korrelasjoner i basalgangliene. Det bør bemerkes at rollene for feillokalisering og endringer i spesifiseringsprosesser ikke ble støttet av genontologi-analyser. Dette illustrerer at videre undersøkelser av BP er nødvendig.

Preface

The work presented in this thesis was conducted at the Department of Biotechnology and Food Sciences at the Norwegian University of Science and Technology (NTNU) under supervision of professor Eivind Almaas. It marks the end of my Master of Science (MSc) degree in Biotechnology, where I have specialized in Systems Biology.

First, I would like to thank my supervisor, professor Eivind Almaas, for his guidance, encouragement and enthusiasm. His positivity and can-do attitude have inspired me throughout my thesis, especially when facing unforeseen difficulties. I would also like to express my gratitude to my co-supervisor, André Voigt. Our weekly meetings, thorough discussions and his excellent feedback have had a major impact and helped to shape the content of my thesis. Furthermore, his assistance with the more technical aspects of this thesis has been indispensable and I thank him for his patience with my endless questions in this regard. I am also grateful for being a part of the Almaas Lab group during my MSc degree. Not only has this allowed me to benefit from the expertise within this group, but also provided a motivating and inspiring work environment.

I also wish to thank my friends and family for their unconditional support and optimism. I really appreciate that they allow me dive into long talks about my project, even when they have no idea what I am talking about. Furthermore, I would like to thank my best friend Andrine, for making sure that I take breaks from working. I have truly appreciated and enjoyed all of our too-long lunches and too-late game nights. Finally, I wish to give a tremendous thank you to my loving and supporting boyfriend Thomas, who has heard more about correlations than any boyfriend should ever have to endure.

Anna Sunniva Hessevik
Trondheim, May 2022

Table of Contents

List of Tables

List of Figures

Abbreviations

BP	Bipolar Disorder
cDNA	complementary DNA
CN	Consensus Network
CoDiNA	Co-expression Differential Network Analysis
CSD	Conserved, Specific, Differentiated
df	Degrees of Freedom
DLPFC	Dorsolateral Prefrontal Cortex
DNA	Deoxyribonucleic Acid
ER	Endoplasmic Reticulum
FDR	False Discovery Rate
FWER	Family-Wise Error Rate
GEO	Gene Expression Omnibus
GO	Gene Ontology
GTEx	Genotype-Tissue Expression
GWAS	Genome-Wide Association Study
iPSC	induced Pluripotent Stem Cells
MAE	Mean Absolute Error
MAS5	Micro Array Suite 5.0
mdDA	Mesodiencephalic Dopaminergic
mRNA	messenger RNA
NSC	Neural Stem Cell
NTNU	Norwegian University of Science and Technology
PCR	Polymerase Chain Reaction
PMI	Post-Mortem Interval
pre-mRNA	precursor mRNA
RMA	Robust Multi-array Analysis
RMSE	Root Mean Square Error
RNA	Ribonucleic Acid
RNA-Seq	RNA Sequencing
SNP	Single Nucleotide Polymorphism
SRP	Signal Recognition Particle
TO	Topological Overlap
wTO	weighted Topological Overlap

Chapter 1

Introduction

The human body consists of trillions of cells [1]. These cells can be categorized into different cell types, which differ enormously in structure and function. Initially, it was believed that cells lost genes as they realized their cellular fates [2, p. 369]. It is now known, however, that the distinctions between cell types are not generally caused by alterations of deoxyribonucleic acid (DNA), which can be seen as a ”cookbook” containing the ”recipes” for the components of all cells. Rather, different cell types arise due to production and buildup of different sets of ribonucleic acid (RNA) and protein molecules [2, p. 369]. RNAs may be thought of as copies of specific ”recipes” in the DNA, while proteins may be seen as the ”dishes” created from these recipes. The flow of genetic information from DNA to RNA and subsequently to proteins is fundamental to molecular biology and has even been termed the central dogma [2, p. 299]. Some genes are expressed in all cells, while others are only expressed in specific types. The level of expression may further contribute to distinctions between cell types. Moreover, gene expression patterns vary during the life time of a cell and are affected by external factors [2, p. 371-372]. Thus, gene expression patterns and the interplay between the expressed components are crucial for proper function of different cell types.

Despite a complex interplay between a cell’s components, molecular biology has traditionally adopted a reductionist thinking when studying cells [3]. This means that cells have typically been reduced to smaller pieces, which then are studied in isolation. Such approaches will not be able to identify *emergent* behaviours of cells, meaning properties that result from the system as whole and cannot be assigned to individual components on their own [4]. As an example, the human genome project was completed in 2001 and produced the first draft of the human genes [5, 6]. However, a list of the human genes is not sufficient for a complete understanding of how cells function and how diseases arise. It is also necessary to investigate how different genes, as well as other cellular components, interact [7]. The large growth in genomics and high-throughput technologies now allow molecular biology to adopt a systems approach [3, 4], giving rise to systems biology.

Complex systems, including biological systems, may be represented and investigated using network approaches [7]. Let us start with a quite straightforward example, such as a social system where it is wishful to examine friendships between individuals. The components in the system, in this example the individuals, are typically referred to as *nodes* or *vertices* when considering the system as a network. Direct interactions between them, such as a friendship between two individuals, are called *links* or *edges*. As a result, the network approach creates a representation of the individuals and their friendships in our social sys-

tem. Similarly, a network approach may be applied to complex biological systems. An example of a biological network is a *protein-protein interaction network*. Here, nodes represent proteins and two proteins are defined as ”friends” if they are capable of binding to each other in a cell [7]. An alternative biological network is called *gene co-expression network*. The nodes in these networks generally represent genes and are ”friends” if they are *co-expressed*, meaning that they are simultaneously expressed (or produced) in a cell [8]. Recently, an extension of the gene co-expression network has been developed and is called *differential gene co-expression network*. Once again, the nodes typically correspond to genes and links are based on their co-expressions. However, this network type focuses on differences in co-expression patterns between two or more conditions [9, 10]. Such networks may be useful for comparison of diseased and healthy samples, as well as other case-control studies [9, 10].

Diseases can arise due to perturbation or breakdown in the molecular network of an individual [7]. As nature is not perfect, errors may be introduced in a cell's genetic material. These errors are known as *mutations*. If the mutations arise in germ cells, and are passed to the offspring, the mutations become inherited [11]. Most mutations have minor impact, and may not even have noticeable effects [11], but some cause disease. Conditions such as albinism, hemophilia and congenital deafness are caused by single mutations. Other diseases, including diabetes and arthritis, arise from an interplay between many different genes and may be strongly affected by environmental factors [2, p. 493]. Differential gene co-expression networks, as well as other systems biology approaches, may be useful for elucidating such complex interplays and may contribute to a better understanding of the mechanisms underlying the diseases.

Bipolar disorder (BP) is an example of a genetically complex disorder, and is affected by genetic as well as environmental factors [12, 13]. This disorder is characterized by large mood swings, which range from emotional highs to lows [12, 13, 14, p. 123-154]. BP confers serious consequences for affected individuals. It often leads to reduced quality of life [12] and has the highest suicide rate among affective disorders [15]. In addition, this disorder affects more than 1% of the world's population [16] and has a typical onset in young adulthood [14, p. 130,136]. Thus, BP also results in high costs to society due to health-care costs and costs of disability [12]. The impact of BP on patients and their family, as well as society and economy, highlights the need for further investigation and understanding. As it has been challenging to diagnose BP due to wide-ranging symptoms, most studies so far have been based on small sample sizes [13]. Consequently, it is of great interest to achieve larger sample sizes to improve our understanding of genetics and altered molecular networks in BP. According to Gordovez and McMahon [13], ”*meta-analysis of multiple independent samples have perhaps the best likelihood of success*” as an approach to increase sample sizes. Even though Gordovez and McMahon [13] presented this claim in the context of next-generation sequencing technologies that searched for BP risk genes, it also seems promising as a method to increase sample size for differential gene co-expression networks.

The aim of this thesis is to investigate differences in gene co-expression patterns between BP and control samples. This will be conducted by using a systematic framework, known as CSD [9], for differential co-expression network analysis. The goal is to identify genes and potential pathways that contribute to BP. As most BP data sets suffer from small sample sizes, a second aim of this thesis is to develop and test methods for combining several data sets into one common, or *consensus*, CSD network. It is believed that this will indicate the most consistent alterations in BP.

Chapter 2

Theoretical Background

This chapter will introduce the theoretical foundations for the methods and data analyses included in this thesis. Some of the topics are directly relevant for the subsequent chapters, while others are included to give a more profound understanding of the underlying concepts. This chapter starts with an introduction of BP, the investigated disorder in this thesis. Next, a general description of networks and some network characteristics are presented. This is followed by a more thorough investigation of specific network types and their underlying measurements. In particular, these network types include (differential) gene co-expression networks, where the CSD approach is in focus, and consensus networks (CNs). Finally, the chapter is ended by a presentation of several statistical methods relevant for comparison and analysis of networks.

2.1 Bipolar Disorder

BP is a collective term for disorders characterized by biphasic patterns expressed through changes in emotions, energy and thoughts of affected individuals. Generally, BP is manifested as phases of mania/hypomania and depression [12], possibly interspaced by periods called euthymia where individuals are free of symptoms [13]. Manic episodes include elevated, expansive or, in some cases, irritable mood. These episodes are also accompanied by other symptoms, such as a feeling of increased energy and self-esteem, reduced need for sleep and/or psychotic symptoms. Hypomania is defined as a milder and shorter version of mania. In contrast, depressive phases typically include depressed mood (a feeling of sadness), loss of interest or pleasure, reduced energy and increased need for sleep. Psychosis may also occur in this phase [12, 14, p. 124-125]. Moreover, mixed features are common in BP and refer to episodes where symptoms from mania/hypomania and depression are manifested at the same time [14, p. 149, 17, p. 72] . In addition to mood episodes, BP may also be associated with cognitive symptoms. This includes changes in reaction time, memory and executive functions [12]. The "*Diagnostic and Statistical Manual of Mental Disorders*" (DSM-5) divides BP into several subcategories depending on the duration and severity of mania/hypomania and depression, as well as the possibility of triggering this disorder by substance/medication abuse or another medical condition [14, p. 123]. BP is also known to be comorbid with other psychiatric and nonpsychiatric disorders [12, 14, p. 132-139].

BP has a complex genetic background. It is among the most heritable psychiatric disorders [12, 13] and has an estimated heritability between 68% and 84% (for both BP I and II) [17,

p. 420]. According to Vieta et al. [12], the best model for this disorder is multifactorial and includes gene-environment interactions. Genetic and environmental factors may give rise to neuronal changes which result in altered circuitry in the brain of individuals with BP [12]. Several genome-wide significant loci have been identified through genome-wide association studies (GWAS) [12, 13]. In general, associated genes have been related to calcium signal transmission, glutamatergic systems, hormone regulation as well as immune and histone pathways [12]. It should be noted that individual associated genes only have a small effect on risk [12].

Several brain areas appear to be affected in BP patients. In general, neuroimaging studies of euthymic BP patients indicate a reduced responsiveness of dorsolateral prefrontal cortex (DLPFC), dorsomedial prefrontal cortex and dorsal anterior cingulate gyrus. These areas are associated with cognitive control [18]. At the same time, neuroimaging studies indicate an increased responsiveness in ventrolateral prefrontal cortex, ventral anterior cingulate gyrus and amygdala. These areas are involved in emotional regulation [18]. Together, this reflects the wide-ranging symptoms observed in BP.

2.2 Network Theory

The world is full of complex systems. These are parts of our everyday life, for instance as social systems. As explained in the introduction (Chapter 1), systems may be represented as networks which display interactions as links/edges between the components, nodes/vertices, of the system. Networks may be used to represent biological systems, such as a network of protein-protein interactions or differential gene co-expression networks. The latter is the focus of this thesis, and will be explained in detail in subsequent sections. This section will introduce some fundamental characteristics of networks.

2.2.1 Basic Network Properties

Networks are composed of nodes which are connected to each other through links. The links are either *undirected* or *directed*, as shown in Figure 2.1. If node A is connected to node B through an undirected link, then B is also connected to A. Directed links however, will only go in one direction. An undirected network consists exclusively of undirected links, otherwise it is defined as a directed network. Differential gene co-expression networks, which will be the main focus in this thesis, can be represented as undirected networks [8]. Thus, this section focuses on undirected networks and their characteristics.

A network can be represented mathematically by an *adjacency matrix,* $A = [a_{ij}]$. The adjacency matrix is a square matrix consisting of N rows and N columns [7]. The values of the elements are determined by equation 2.1. This equation describes the adjacency matrix of an *unweighted* network, where the value of each element is either zero or one. For a *weighted* network, a_{ij} is equal to a weight, w_{ij}, if there is a link between i and j.

$$a_{ij} = \begin{cases} 1, & \text{if there is a link from } i \text{ to } j \\ 0, & \text{otherwise} \end{cases} \tag{2.1}$$

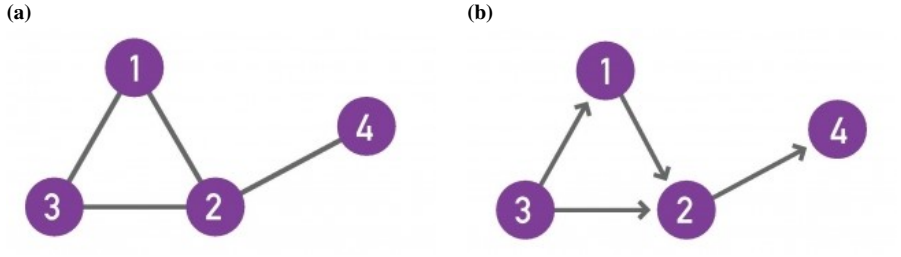


Figure 2.1: (a) Undirected and (b) directed networks with four nodes and four links. From [7], CC BY.

In an undirected network, the adjacency matrix will be symmetric. This means that $a_{ij} = a_{ji}$ [7]. The relationship between the adjacency matrix and the graphical representation of the network is illustrated in Figure 2.2.

The *degree*, k_i, is the number of links between a node i and its neighbours [7]. Using Figure 2.2 as an example, node 2 has a degree of 3. Nodes with high degrees are referred to as *hubs*. The *average degree*, $\langle k \rangle$, in an undirected network is given by equation 2.2.

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i = \frac{2L}{N} \tag{2.2}$$

where N represents number of nodes, k_i represents degree of node i and L represents total number of links.

The described network properties focus on direct interactions between two nodes. The *clustering coefficient,* C_i, on the other hand, measures the extent that the neighbours of a given node link to each other [7]. The *average clustering coefficient*, $\langle C_i \rangle$, indicates the tendency for formation of groups in the network [19]. The definition of C_i for an unweighted, undirected network is given in equation 2.3. In this case, the clustering coef-

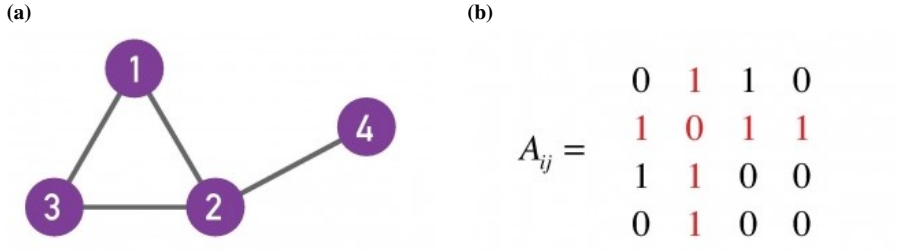


Figure 2.2: (a) An undirected network and (b) its corresponding adjacency matrix. The relationships of node 2 are highlighted in red in the adjacency matrix and indicate that this node is connected to all other nodes in the network. From [7], CC BY.

ficient of node i is equal to one if all neighbours of node i link to each other.

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

where k_i represents number of neighbours of node i and L_i represents number of links between these neighbours.

Topological overlap (TO) is similar to the clustering coefficient, but focuses on shared neighbours between two nodes [20]. The definition of TO in an unweighted network is [8]:

$$\omega_{ij}^{TO} = \frac{\sum_u a_{iu}a_{uj} + a_{ij}}{min(k_i, k_j) + 1 - a_{ij}} \tag{2.4}$$

where i and j correspond to two nodes, and k_i and k_j denote the degrees of these nodes. a corresponds to the adjacency matrix and its subscripts denote the specific element. For a weighted network, this measure is called *weighted topological overlap (wTO)* [21, 22] and is defined as:

$$\omega_{ij}^{wTO} = \frac{\sum_u w_{iu}w_{uj} + w_{ij}}{min(K_i, K_j) + 1 - |w_{ij}|} \tag{2.5}$$

where w_{ij} denotes the weight of the link specified by its subscript and K_i corresponds to the weighted connectivity of the node i. The weighted connectivity of node i is defined in equation 2.6 [21].

$$K_i = \sum_{j=1}^{N} |w_{ij}| \tag{2.6}$$

It is noteworthy that the given equation for wTO, equation 2.5, is different from the wTO formula from Zhang and Horvath [8]. The modification allows the elements in the underlying adjacency matrix to take values between -1 and 1 [21]. Thus, wTO may be calculated from an adjacency matrix where the elements correspond to correlation coefficients as any correlation must fall on this interval.

Figure 2.3 illustrates calculation of TO from an undirected, unweighted network. In general, TO is equal to one if the node with the fewest connections in the underlying network, lets say node i, is connected to node j and all neighbours of node i are also neighbours of node j. An example is provided by node A and C in Figure 2.3. If node i and j are not linked to each other, nor share any neighbours, TO is equal to zero [20]. Originally, TO and wTO were used to identify modules [8, 20], see Section 2.2.4. However, it may also be used as an alternative to correlations to represent the similarity between two genes [21–24], see Section 2.3.

To summarize, a network consists of nodes and links, and may be represented by an adjacency matrix. Each link is either defined as directed or undirected. In addition, it may have an associated weight. Furthermore, each node is characterized by several properties, including degree, clustering coefficient and topological overlap.

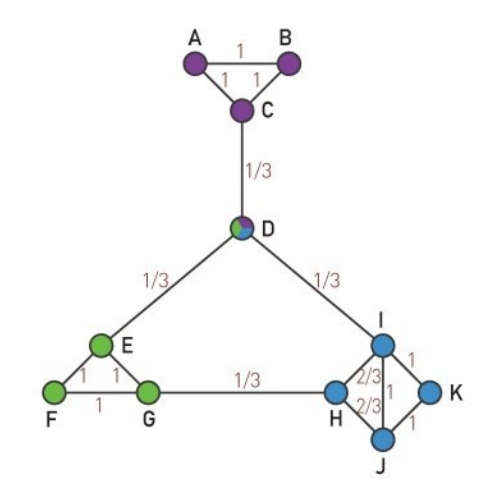


Figure 2.3: Topological overlap (TO) in an unweighted and undirected network consisting of 11 nodes. The numbers on the links correspond to TO calculated from equation 2.4. From [7], CC BY.

2.2.2 Degree Distribution and Scale-Free Networks

The probability that a randomly chosen node has degree k, is represented by the *degree distribution,* p_k, of the network [7]. p_k is given by equation 2.7. The degree distribution is an important network characteristic and is involved in calculations of several other network properties [7].

$$p_k = \frac{N_k}{N} \tag{2.7}$$

where N_k is number of nodes with degree k and N is total number of nodes.

In some random network models, including the Erdős-Rényi model [25, 26], networks are generated by randomly connecting nodes with a probability P. These networks are characterized by a binomial degree distribution, which is typically well approximated by a Poisson distribution [7]. Consequently, this degree distribution is characterized by a peak around the average degree, $\langle k \rangle$, and is independent of network size [7].

In contrast, most biological networks have *scale-free* degree distributions [19]. A comparison of a Poisson and a scale-free degree distribution is presented in Figure 2.4. For scale-free networks, p_k follows, or at least approximates, a power law: $p_k \sim k^{-\gamma}$. The exponent γ is called *degree exponent* and the typical value of this exponent is between 2 and 3 [7]. For these networks, $\langle k \rangle$ is finite even when the number of nodes approaches infinity. However, the second moment, $\langle k^2 \rangle$, approaches infinity under the same conditions. The second moment is involved in the calculation of the variance of the degree distribution. As a result, the variance will approach infinity when the second moment approaches infinity. Consequently, there is no internal "scale" in scale-free networks [7]. The nature of the power law degree distribution thus allows co-existence of many small-degree nodes and

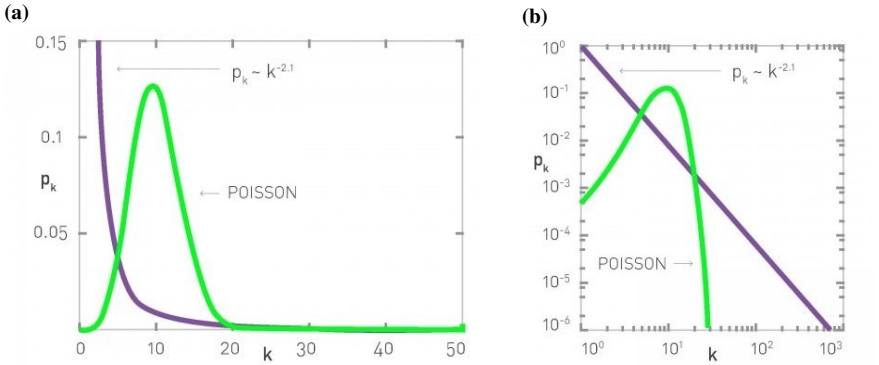


Figure 2.4: (a) Linear and (b) log-log plot of two networks with Poisson and scale-free degree distributions where $\langle k \rangle = 11$ for both networks and $\gamma = 2.1$ for the scale-free network. The scale-free distribution enables the presence of small-degree nodes and hubs. From [7], CC BY.

hubs in real biological networks [7].

Alternative random network models have been developed to generate random networks with scale-free, or other, degree distributions. One example is provided by the the configuration model which generates networks according to a predefined degree sequence [27]. In this model, all nodes are first assigned specific degrees, which can be represented as ”stubs” or ”half-links” as in Figure 2.5. Next, a pair of stubs is randomly selected and connected to each other. This is repeated until all stubs are paired [27]. Consequently, this model allows the degree distribution of a real network to be used as starting point for a random network. A visualization of the configuration model is provided in Figure 2.5.

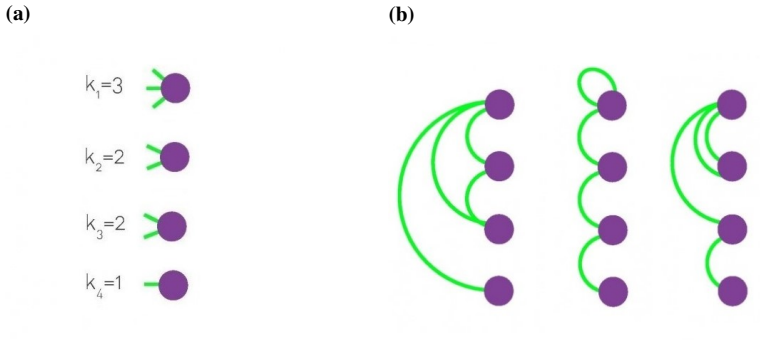


Figure 2.5: The configuration model generates networks from a predefined degree sequence. (a) All nodes are first given specific degrees from the input degree sequence, which are represented as stubs. The stubs are subsequently connected at random which may give rise to the exemplified networks in (b). From [7], CC BY.

To sum up, some random networks are characterized by binomial degree distributions. Real networks on the other hand, have typically scale-free degree distributions which enable co-existence of small-degree nodes and hubs. Some random networks models, such as the configuration model, allow creation of networks with degree distributions which mimic real networks.

2.2.3 Assortativity and Disassortativity

In some networks, nodes with similar degrees tend to link together. This means that hubs generally connect to other hubs and tend to have fewer connections to small-degree nodes. The small-degree nodes on the other hand, tend to link to each other. These networks are known as *assortative* networks [7]. Other networks exhibit the opposite trend, and are known as *disassortative* networks. These exhibit a hub-and-spoke topology, meaning that hubs mainly connect to small-degree nodes [7]. Networks that are not assortative nor disassortative are called *neutral* networks. The assortativity of a network can be measured by r [28], calculated as shown in equation 2.8, which ranges from -1 to 1. Negative values of r indicate disassortativity, while positive values indicate assortativity [28].

$$r = \frac{\sum_{xy} xy(e_{xy} - a_x b_y)}{\sigma_a \sigma_b} \tag{2.8}$$

where a_x and b_y are fractions of links that start and end at nodes with degree x and y, respectively. e_{xy} is the fraction of all links that connect nodes with degree x and y. σ_a and σ_b represent the standard deviations of the distributions a_x and b_y, respectively. Note that r corresponds to the standard Pearson correlation coefficient (see Section 2.3).

2.2.4 Communities

Many networks contain groups of nodes which form dense subgraphs. An obvious example is provided by social networks, where groups of friends cluster together. These groups are called *communities*, or *modules*, and consist of nodes which are more likely to be connected to each other than to other nodes in the network. There exist three different community definitions: cliques, strong communities and weak communities. A *clique* is a complete subgraph of a network, meaning that all nodes in the subgraph connect to each other [7]. This is the strictest definition of a community. In a strong community, each node has more links to nodes within the community than to other nodes [7]. This requirement is expressed as:

$$k_i^{int}(C) > k_i^{ext}(C)$$

where k_i^{int} represents the number of links between node i and other nodes in the community C. k_i^{ext} represents number of links between node i and nodes outside community C.

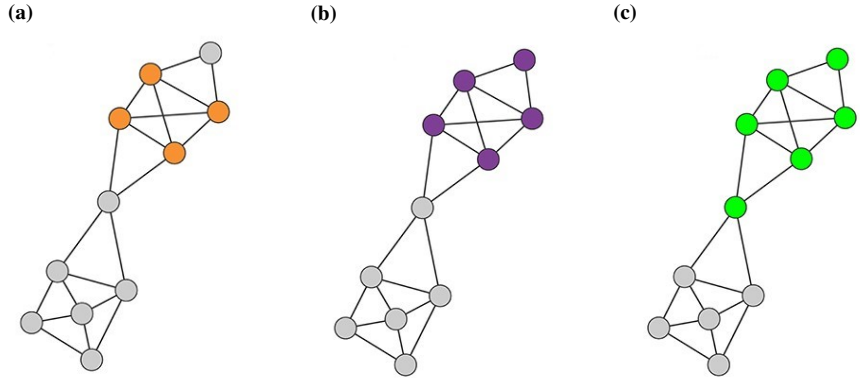


Figure 2.6: Communities can be defined as (a) cliques, (b) strong communities or (c) weak communities. The coloured nodes correspond to nodes which are members of exemplified communities. Additional communities are also present in the networks. From [7], CC BY.

In a weak community, the sum of links within the community must exceed the sum of links between the nodes in the community and nodes outside the community [7]. This is more succinctly stated as:

$$\sum_{i \in C} k_i^{int}(C) > \sum_{i \in C} k_i^{ext}(C)$$

where k_i^{int}, k_i^{ext} and C are defined as above.

Note that all cliques are strong communities and that all strong communities also fulfill the requirements for weak communities. A visual comparison of different community definitions is presented in Figure 2.6.

Brute-force approaches for detection of communities are computationally infeasible due to the large number of possible partition of a network [7]. Hence, several algorithms have been developed to aid the identification of communities in a network. One example is the Louvain algorithm [29]. This method consists of two phases which are repeated iteratively. Initially, all nodes in the network are assigned to different communities. Next, the effect of moving node i from its own community to the communities of its neighbours is evaluated. The estimated effect is based on alteration of *modularity*, which measures the quality of a partition [29]. Node i is then placed in the community which gave the maximum positive modularity gain. If no movement of node i gives a positive gain, i will remain in its own community. This process is repeated for all nodes in the network. Phase one stops when no individual move gives a modularity gain. Note that this implies that a node may be considered several times [29]. The second phase allows construction of a meta-network, meaning a network where nodes correspond to the communities identified in phase one. The weight of the link between two community-nodes in the meta-network is equal to the sum of the weight of links between the original nodes in the corresponding communities

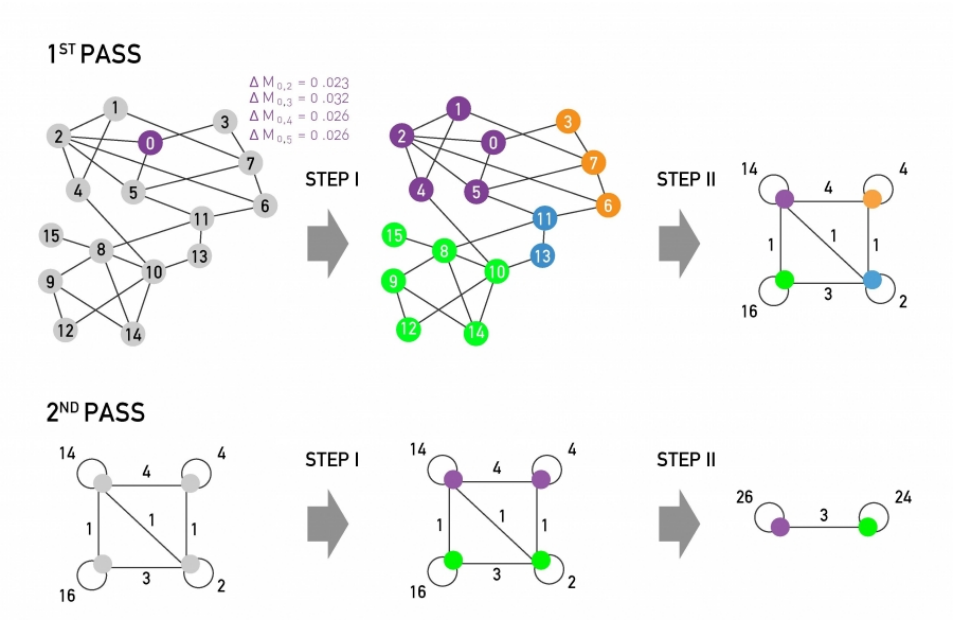


Figure 2.7: The Louvain algorithm may be used to identify communities in a network. It consists of two phases, which are repeated iteratively until there is no gain in modularity. Phase one is based on optimization of modularity by local changes. Phase two aggregates communities to construct a meta-network. From [7] based on [29], CC BY.

[29]. In the next step, phase one may be reapplied to construct communities of communities. This allows investigation of several organization levels in the original network [29]. The entire process is repeated until there is no more changes in the final network [29]. A visualization of the Louvain algorithm is provided in Figure 2.7.

In summary, networks may consist of subgraphs called communities or modules. These are defined as cliques, strong communities or weak communities. The community structure may be examined by the Louvain algorithm.

2.3 Correlations

The human body consists of many different cell types. The genome in all of the somatic cells is the same, but only a subset of the genes are expressed. This subset varies between cell types and with the environment of the cells [2, p. 503]. To evaluate the level of concordance between gene expression patterns of two genes, a *similarity measure* needs to be introduced [8]. All similarity measures quantify the dependence between two sequences of measurements [30]. Examples of similarity measures include wTO [21–24], which was introduced in Section 2.2.1, and correlation [8, 9, 30]. The latter is the focus of this section.

Pairwise correlations, such as *Pearson correlation* and *Spearman rank correlation*, are fre-

quently used as similarity measures [8, 9, 30]. The values of correlations coefficients range from -1 to 1. Positive correlation indicates that the value of the first variable tends to increase as the second variable increases, while a negative correlation represents the inverse relationship. The absolute value of the correlation coefficient indicates the strength of the relationship [31, 32, p. 160]. It is important to note that a statistical significant correlation does not correspond to a causal relationship.

The Pearson correlation coefficient is a measure of a linear relationship between two sets of data [31, 33]. Both of the underlying variables should be continuous and normally distributed. This measure is affected by extreme values and may therefore be strongly influenced by outliers in the data sets [33]. The equation for the Pearson correlation coefficient is given in equation 2.9.

$$\rho_{ij} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{2.9}$$

where X and Y represent two different variables, n corresponds to the number of data points and $\bar{X}$ and $\bar{Y}$ represent the average values for variable X and Y, respectively.

The Spearman rank correlation coefficient can be calculated by inserting the ranks of variable X and Y, instead of raw data, in equation 2.9 [34]. When the sequence of ranks is identical for the two variables, the Spearman rank correlation coefficient takes a value of 1. When the ranks of X vary inversely with the ranks of Y, the Spearman rank correlation coefficient is equal to -1 [34].

In contrast to the Pearson correlation coefficient, the Spearman rank correlation coefficient measures statistical dependency of monotonic, non-linear relationships [34]. This is illustrated in Figure 2.8. In general, the Spearman rank correlation coefficients are more

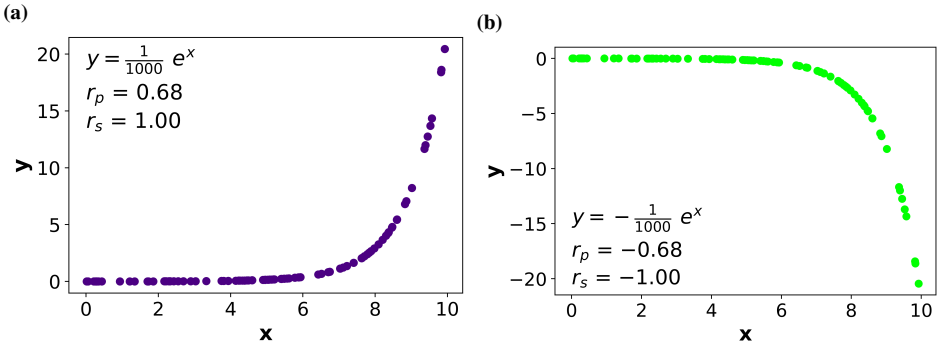


Figure 2.8: Correlation coefficients of monotonically (a) increasing and (b) decreasing relationships where $Y = \pm\frac{1}{1000}e^X$. A monotonic relationship results in perfect Spearman rank correlation coefficient (r_s) even though the dependency is non-linear. Note that the Pearson correlation coefficients (r_p) are not perfect.

robust to outliers than Pearson correlation coefficients [33, 34]. Furthermore, it is applicable to analysis of ordinal and/or non-normally distributed variables [33, 34]. These properties are caused by the non-parametric nature of Spearman rank correlation coefficients [34].

Taken together, correlations may be used as similarity measures to evaluate concordance between gene expression patterns. Pearson correlation is suitable to determine similarity between patterns that are linearly related, while Spearman rank correlation measures dependency of monotonic, non-linear relationships.

2.4 Gene Expression Analysis

The gene expression levels in cells and tissues can be investigated by two main approaches: microarrays and RNA Sequencing (RNA-Seq) [2, p. 503]. Microarrays consist of an array of short DNA sequences, called probes, attached to a microchip. Isolated transcripts from the cells or tissue under investigation are converted to complementary DNA (cDNA) and labelled with fluorescence. Next, the cDNAs are added to the microchip and allowed to hybridize with complementary probes. The labelling of the cDNAs enables estimation of expression levels [2, p. 503]. However, microarrays may require complicated normalization methods to allow comparison of different experiments [35]. This step reduces systematic errors which may arise due to labelling, hybridization and scanning [36]. It exists several approaches for normalization of microarrays, including Micro Array Suite 5.0 (MAS5) [37] and Robust Multi-array Analysis (RMA) [38].

RNA-Seq sequences all RNAs present in a cell/tissue. First, the RNAs in the sample are converted to a cDNA library which may be amplified. The cDNA library is subsequently sequenced. The resulting reads are identified and level of expression can be calculated [35]. Initially, Wang et al. [35] optimistically claimed that RNA-Seq *"can capture transcriptome dynamics across different tissues or conditions without sophisticated normalization of data sets"*. However, Robinson and Oshlack [39] illustrated that normalization methods are often an important requirement. The normalization methods may account for within-sample or between-sample effects. Within-sample effects refer to factors that influence comparison of read counts between genes within one sample [40], such as gene length [41] and GC-content [42]. Between-sample effects, such as sequencing depth [39], describe factors that influence comparison of read counts for individual genes between samples [40].

2.5 Gene Co-Expression Networks

Expression profiles from a gene expression analysis may be used to analyse co-expression of genes. Functionally related genes are likely to be regulated in a similar way in different conditions. Hence, gene functions may be deduced from these analyses [2, p. 504, 43]. This has been used to identify pathways associated with specific disorders, including autism spectrum disorders [44]. However, it is important to note that genes which share a

regulatory DNA motif by chance may be expressed in the same conditions as well. As a result, co-expression does not automatically indicate a functional relationship [9, 43].

Gene co-expressions can be represented and analysed as an undirected network. The nodes in this network correspond to genes, while the links correspond to (significant) co-expressions [8]. A gene co-expression network can be created by the following steps [8]:

1. Collect gene expression data from experiments, for instance from microarrays or RNA-Seq.
2. Define a measure of similarity, typically pairwise correlation between gene expressions.
3. Transform the similarity matrix to an adjacency matrix. All elements along the diagonal in the adjacency matrix are set to zero. The next step depends on whether the resulting co-expression network is unweighted or weighted:
   (a) Unweighted gene co-expression network: introduce a ”hard” threshold. All values in the similarity matrix with a value above the threshold will attain a value of one in the adjacency matrix. The remaining values will be set to zero.
   (b) Weighted gene co-expression network: introduce a ”soft” threshold. The soft threshold converts the similarity score to a connection weight.

Microarrays and RNA-Seq are often used to compare gene expression profiles between different cell types or cells subjected to different conditions. Thus, differences in gene co-expressions can be evaluated [2, p. 503-504]. *Differential gene co-expression networks* may be used to identify and analyse these differences. Several of the differential gene co-expression network methods construct separate gene co-expression networks for each of the studied conditions. The network structures are subsequently compared to investigate potential differences in the underlying systems. Simple comparisons include evaluation of alterations in node degrees and identification of links which are related to a specific condition [9, 10]. Other differential gene co-expression network methods construct one common network which represents the differences between the investigated conditions. Each link in the network is given a score based on the alterations in gene co-expression in the different conditions [9]. Different methods use different strategies to define this score. The resulting links are subsequently filtered to exclude non-significant changes [9].

In sum, gene co-expression networks aid analysis of expression profiles from cells/tissue on a system level. Differential gene co-expression networks facilitate the identification of differences, and similarities, between different gene expression profiles. Several methods have been develop to assist these analyses.

2.6 CSD Analysis

The CSD approach was developed at NTNU by Voigt et al. [9] and aids the analysis of differential gene co-expression networks. In this method, co-expressions are defined as conserved (C), specific (S) or differentiated (D). Conserved co-expression means that a gene pair is similarly correlated in two condition, while specific co-expression implies that two genes are only correlated in one condition. Differentiated co-expression indicates that a gene pair is negatively correlated in one condition and positively correlated in another [9]. This section will describe the basic steps of the CSD approach.

The CSD approach is based on the identification of pairwise gene co-expression scores $\rho_{ij,k}$, where *i* and *j* denotes two genes and *k* denotes the condition. The calculated scores correspond to Spearman rank correlation coefficients. The variation in co-expression within a given condition is measured by the standard error of the mean, $\sigma_{ij,k}$, calculated form a set of Spearman rank correlation coefficients from independent subsamples from the given condition. A detailed description of how to select subsamples is given by Voigt et al. [9]. A minimum of 49 data points for each condition is required to obtain a satisfying accuracy for the standard error of $\rho_{ij,k}$ [9]. Furthermore, Voigt et al. [9] argue that a minimum of 49 data points will reduce the impact of stochastic effects related to small sample sizes [9]. In case of inadequate sample size, it is recommended to omit the subsampling [9].

Conserved, specific and differentiated co-expressions are associated with three pairwise comparative scores, which are given in equation 2.10, 2.11 and 2.12. A visual representation of the classification of gene co-expressions as conserved, specific and differentiated is provided in Figure 2.9.

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

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

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

where $\rho_{ij,k}$ and $\sigma_{ij,k}$ represent Spearman rank correlation coefficient and the standard error of the mean for the co-expression of gene i and j in condition k, respectively.

C_{ij}, S_{ij} and D_{ij} are positive values and range from zero to infinity. However, these are not directly comparable with each other. Consequently, a new value, X_p, must be introduced. This value maps C_{ij}, S_{ij} and D_{ij} to a common scale, thus enabling the combination of C-, S- and D-links into a single network using a common value of p [9]. In the resulting CSD network, links are treated as unweighted. Note that p is not a significance threshold in the calculations, but is referred to as the *importance level* [9]. X_p represents a threshold value and its definition is given in equation 2.13.

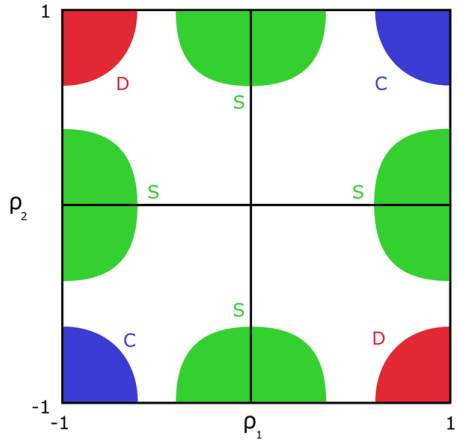


Figure 2.9: The CSD approach categorizes gene co-expressions from two different conditions as conserved (C), specific (S) or differentiated (D). This is based on their respective Spearman rank correlation coefficient, ρ_1 and ρ_2. If a gene co-expression exhibits a large value for C, S or D, it will be located in the correspondingly coloured area in this plot (blue, green or red, respectively). From [9], CC BY.

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

where X is a random variable from the underlying distribution of either C-, S- or D-links, and m corresponds to the number of samples s_i, with size L, drawn from a data set of M points. p is defined as $1/L$ where $L << M$.

Links in the final CSD network are categorized as C, S or D. A node however, may be connected to other nodes by either of these link types. A *homogeneity measure*, given in equation 2.14, is introduced to evaluate the identities and number of links between node i and its neighbours. This measure may take values from $1/3$ to 1. The lowest value $(1/3)$ indicates equal number of C, S and D links for the given node, while a value of 1 indicates the presence of just one type of link from said node.

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

where $k_{j,i}$ corresponds to the number of j-type links (either C, S or D) between node i and its neighbours. k_i denotes total degree of node i.

In conclusion, the CSD approach classifies links in a differential gene co-expression networks as conserved, specific or differentiated. The links may be combined into a common network by defining an importance level and calculating the threshold value for each link type. Nodes in the network may be given a homogeneity value, which indicates if a node is mainly connected to its neighbours by one or several types of connections.

2.7 Consensus Network

Most network studies rely on data from one data set. However, when networks based on similar data sets are compared, they often show considerable variation [23]. These differences may be due to biological or technical variation. Consequently, Berto et al. [23] and Gysi et al. [22] have developed methods for constructing a consensus network (CN) from multiple data sets. According to Gysi et al. [22], a CN should have higher biological confidence than its input networks as the CN focuses on the commonalities across the input networks. In this section, some potential methods for constructing CNs will be presented.

Current methods by Berto et al. [23] and Gysi et al. [22] for constructing consensus gene co-expression networks are based on wTO. It has been argued that the use of wTO in a network setting reduces the effect of false positives compared to networks based on correlation coefficients [21]. Voigt and Almaas [24] also demonstrated that wTO may increase the fidelity of gene co-expression networks, particularly when a data set contains few samples. However, it is noteworthy that the topology of gene co-expression networks based on wTO was found to be quite different from networks based on correlations [24]. Furthermore, the use of CNs based on wTO is limited for subsequent CSD analysis as this method is based on Spearman rank correlation coefficients [9]. To my knowledge, no method for combining gene co-expression networks or differential gene co-expression networks based on correlation coefficients exists. However, some approaches have been suggested for estimating mean correlation coefficients. These might be promising for the development of a new method for constructing consensus CSD networks. The remainder of this section focuses on these procedures.

Fisher [45, p. 199], Hedges and Olkin [46, p. 231] and Rosenthal [47, p. 73] argued that the correlation coefficients should be transformed using *Fisher's Z transformation* before estimating the mean of the correlation coefficients. This transformation is shown in equation 2.15 [45, p. 200].

$$Z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) \tag{2.15}$$

where r is the estimated correlation coefficient. The same transformation may be used when r corresponds specifically to the Spearman rank correlation coefficient [34].

The transformation of a correlation coefficient to Z has three main advantages, which are also reflected in Figure 2.10:

1. The standard error of Z may be regarded as independent of the true value of the correlation in the population [45, p. 200].

2. The Zs are approximately normally distributed. In contrast, the typical distribution of estimated correlation coefficients is far from normal, especially for high values [45, p. 201].

3. The form of the distribution of Zs is nearly constant. In contrast, the distribution of

estimated correlation coefficients changes form if the true value of the correlation is altered [45, p. 201].

If a fixed-effect model is assumed, meaning that the true correlation coefficient in the population is assumed to be constant and equal for all included studies [48], the Fisher's Zs may be averaged as shown in equation 2.16 [47, p. 74].

$$\bar{Z} = \frac{\sum_{i=1}^{k}(n_i - 3)Z_i}{\sum_{i=1}^{k}(n_i - 3)} \tag{2.16}$$

where Z_i is the Fisher's Z transformed correlation coefficient in study i and n_i is the sample size of study i.

The Z value can be backtransformed to a correlation coefficient as shown in equation 2.17 [47, p. 71].

$$r = \frac{e^{2Z} - 1}{e^{2Z} + 1} \tag{2.17}$$

According to Fisher [45, p. 207], there is a small bias when averaging correlation coefficients in this way. The averaged Z is slightly larger than the value of the true population parameter that it estimates. Hence, the estimated correlation coefficient is somewhat exaggerated. This is reflected in Figure 2.10b where the ordinate of zero error is not centrally placed when the true correlation coefficient is 0.8. Furthermore, Sheppard's adjustment (a correction for grouping errors) is omitted in the calculation of correlation coefficients when calculating Fisher's Zs [45, p. 207]. This results in a second systematic error, but in the opposite direction of that described above. However, this second systematic error is typically small [45, p. 208].

Even though the Fisher's Z transformation is often used to estimate mean correlation coefficients, Schmidt and Hunter [48] are critical to this method. This is due to the bias described above. According to Schmidt and Hunter [48], the positive bias introduced by Fisher's Z transformation "*is always greater in absolute value than the bias in the untransformed correlation*". Hence, Schmidt and Hunter [49] suggest to calculate the weighted mean from the untransformed correlation coefficients. This is shown in equation 2.18. The method by Schmidt and Hunter [49] also offers the opportunity to correct for study design artifacts, for instance sampling error. However, this requires information about the size and nature of the artifacts from the included studies [49]. Consequently, the description and subsequent use of the Hunter-Schmidt method have been limited to its simplest version in this thesis.

$$\bar{r} = \frac{\sum n_i r_i}{\sum n_i} \tag{2.18}$$

where n_i indicates the number of samples in study i and r_i represents the estimated correlation coefficient in study i.

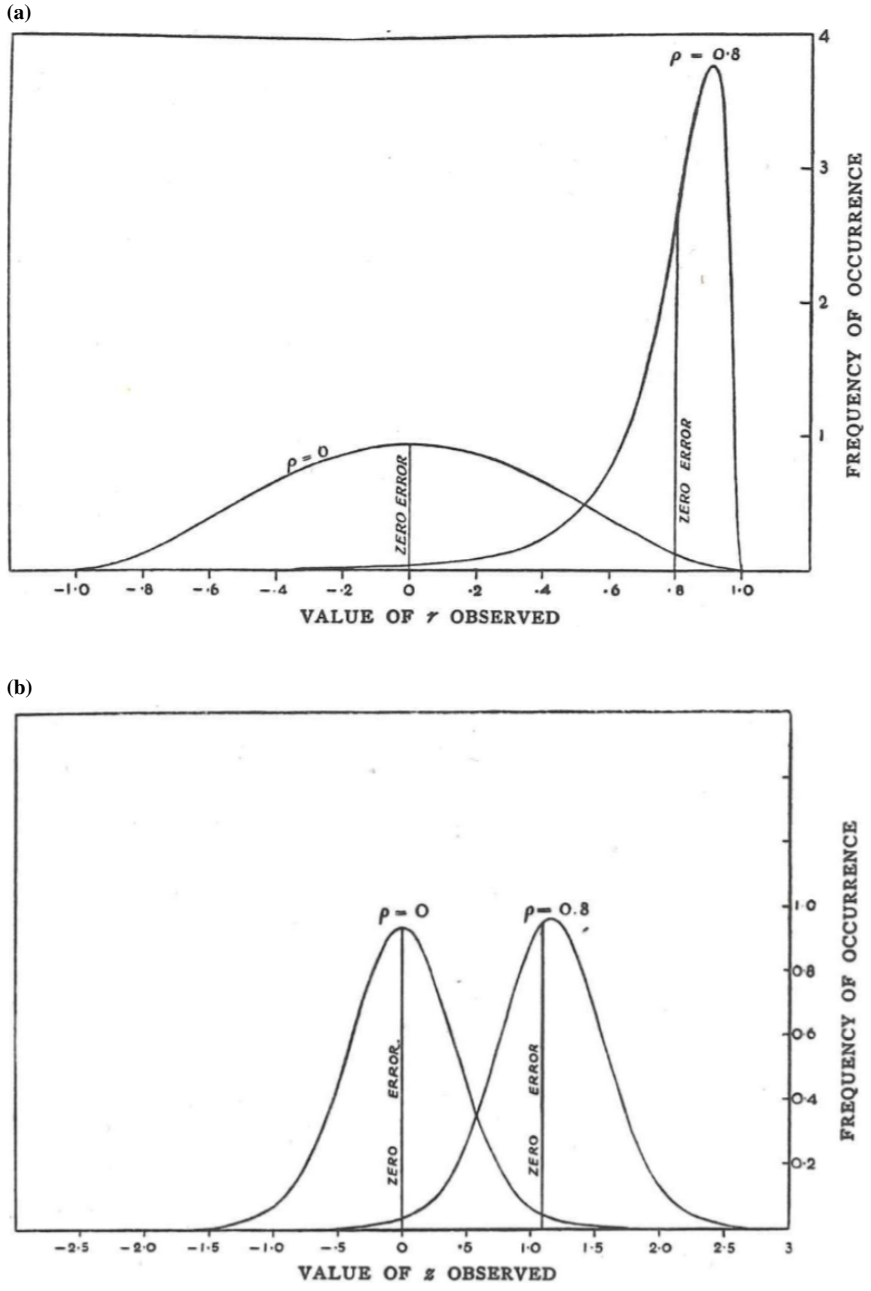


Figure 2.10: (a) Distributions of correlation coefficients from eight observations where the true correlations are 0 and 0.8. (b) Distribution of Zs from the Fisher's Z transformed values from (a). Both curves in (b) are nearly normally distributed and have almost equal heights. From [45, p. 202].

Schmidt and Hunter [48] do however acknowledge that the Fisher's Z transformation may be useful in some cases. As this transformation allows the standard error to only depend on sample size [45, p. 200], and thus to be independent of the true value of the correlation coefficient, it is useful in statistical tests [48]. Nevertheless, Schmidt and Hunter [49] argue that the weighted untransformed average of observed correlations is the best estimate of mean correlation.

As described, Schmidt and Hunter [48] are reluctant to the use of Fisher's Z transformation when estimating mean correlation coefficients. At the same time, Hedges and Olkin [46, p. 230] are critical to the use of linear combinations of untransformed correlation coefficients for this calculation. They claimed that the sample sizes of the included studies have to be quite large in order to recommend the use of this method [46, p. 230]. Taken together, both Fisher's Z transformed and weighted untransformed averages of correlation coefficients seem to be possible methods for estimating combined correlation coefficients and may be useful for creating consensus CSD networks. However, both methods have limitations and lead to bias in the final estimate.

2.8 Gene Ontology

In the study of biological networks, it is often beneficial to describe and group genes and gene products using Gene Ontology (GO). In general, an *ontology* formalizes knowledge about concepts by describing and classifying them in relation to each other [50]. Each ontology term is provided with a name, synonyms, definition and a unique ID [50]. GO is a specific ontology related to biological processes, molecular functions and cell components [50]. Furthermore, GO facilitates the representation of biological information in a computer-friendly way, enables connections across different biological databases and eases the analysis of large data sets [50]. The latter is especially useful for (differential) gene co-expression networks, as groups of genes may be analysed to investigate if some GO terms are over- or underrepresented in the network [50]. As a result, this may give clues about biological functions, roles or locations of a group of genes/gene products.

2.9 Statistics

The use of statistics is necessary to analyse networks and compare potential methods for constructing CNs. This section gives a brief introduction to statistical methods that are relevant for this thesis, including the Jaccard index, root mean square error (RMSE), hypothesis testing and the multiple comparison problem.

2.9.1 Jaccard Index

The similarity between two sample sets may be reflected by the *Jaccard index*, also known as *Jaccard similarity coefficient*, *Tanimoto index* or *Tanimoto coefficient*. This index is defined as the size of the intersection divided by the size of the union, as shown in equation 2.19 [51, 52]. A visualization of the intersection and union is provided in Figure 2.11. The

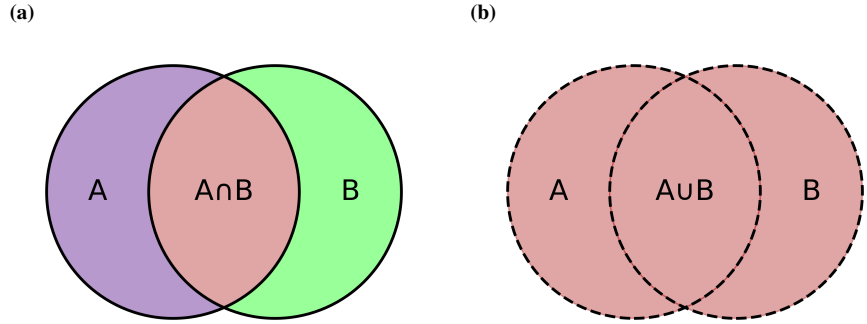


Figure 2.11: The Jaccard index is defined as the size of (a) the intersection divided by the size of (b) the union of two sample sets.

Jaccard index is a useful measure as it allows quantification of overlap between two sample sets [52]. Hence, this statistic is used in many disciplines, including ecology [51–53], machine learning [54] and biological network analyses [24].

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{2.19}$$

where A and B are two sample sets, and $J(A,B) = 1$ if $|A \cup B| = 0$.

2.9.2 Root Mean Square Error

The root mean square error (RMSE) is used to measure model performance [55]. This metric is utilized in several disciplines, including climate research studies [56], biomedicine [57] and epidemiology [58]. Its definition is given in equation 2.20, where it is assumed that the errors are unbiased and follow a normal distribution [55]. In general, a lower RMSE indicates a better fit between the observed and predicted values.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{2.20}$$

where $\hat{y}_i$ is the predicted value, y_i is the observed value and n is number of observations.

Other metrics, such as mean absolute error (MAE), may be used as alternatives to RMSE for evaluating model performance. Importantly, MAE gives equal weight to all errors. RMSE on the other hand, gives more weight to errors with large absolute values [55]. Consequently, RMSE is more sensitive to outliers. According to Chain and Draxler [55], there is not an agreement for which metric to use as a standard when evaluating model performance.

2.9.3 Hypothesis testing

Hypothesis testing is performed to test two hypotheses against each other. The first hypothesis is called a *null hypothesis,* H_0, while the other hypothesis is called an *alternative hypothesis,* H_1. The purpose of hypothesis testing is to determine if H_0 can be rejected [32, p. 231]. In general, there are two main ways of formulating the hypotheses of statistical tests. One-tailed tests examine if the estimated value is significantly different from a reference value in one direction, either larger or smaller. Two-tailed tests on the other hand, examine if the estimated and reference values are significantly different in either direction [32, p. 257].

There are two possible erroneous conclusions that can occur in hypothesis testing: type I and type II errors [32, p. 232]. Type I errors, also called *false positives*, refer to rejection of H_0 even though H_0 is correct. Type II errors, also called *false negatives*, refer to keeping of H_0 even though H_0 is false. It is wishful to avoid both type I and type II errors in hypothesis testing. However, there is a trade-off between them. If the goal is to minimize type I errors, this will generally increase the occurrence of type II errors. The reverse is also true [32, p. 232]. In general, hypothesis testing focuses predominantly on type I errors.

When hypothesis testing is performed, it is necessary to define a *significance level*. This level represents the acceptable level of making type I errors [32, p. 233]. During the hypothesis testing, a *P value* is calculated. This value represents the probability of sampling a test statistic which is at least as extreme as the observed result, given that H_0 is true [32, p. 268]. H_0 is rejected if the *P* value is lower than the significance level. An overview of relevant statistical tests for this thesis is presented in Table 2.1.

In conclusion, hypothesis testing is used to test an alternative hypothesis against a null hypothesis. This typically involves the calculation of a *P* value, which helps to determine if the null hypothesis should be rejected.

Table 2.1: Summary of relevant statistical tests.

Test	Description	Reference(s)
One sample t-test	Tests if the population mean is statistically different from a given value. It is assumed that the measurements are normally distributed.	[32, p. 271].
Independent samples t-test	Tests if there is a statistical difference between the means of two (unpaired) groups. It is assumed that the underlying measurements of both groups are independent of each other and normally distributed.	[32, p. 340].
Wilcoxon signed-rank test	Non-parametric test to investigate if medians of two paired groups are statistically different. It is assumed that the distribution of the differences between the groups are continuous and symmetrical, but normally distributed data is not required.	[32, p. 360]
Binomial test	Tests if the number of observed successes is different from a hypothesized probability of success. Binomial data is required and it is assumed that the probability of success is the same in all experiments and that the experiments are independent.	[32, p. 360, 59, p. 88-89]
Fisher's exact test	Tests if one binary variable is related to another. A hypergeometric distribution of the data is assumed. This test is used by many tools to investigate enrichment of Gene Ontology terms.	[59, p. 355-359, 60–62]

2.9.4 Multiple Comparison Problem

The creation of (differential) gene co-expression networks demands particular attention as thousands of statistical tests are performed simultaneously in one analysis. This increases the probability of finding a significant result due to chance if the conventional thresholds for *P* values are used [63]. This problem is called the *multiple comparison problem*. Typical correction methods aim to control the type I error rate when multiple tests are performed simultaneously [63]. These methods include the *Bonferroni correction* and the *Benjamini-Hochberg correction* presented below.

The Bonferroni correction controls the probability of making at least one type I error, which is known as the family-wise error rate (FWER) [63]. In this method, a new threshold is established by dividing the significance level by the total number of statistical tests performed in the analysis. H_0 is rejected if the *P* value is lower than the new threshold [63].

The Benjamini-Hochberg correction controls the false discovery rate (FDR), which is defined as the expected proportion of false positives among positive tests [64]. If several hypotheses are tested at the same time and all null hypotheses are true, then FDR is equal to FWER. However, if only some of the null hypotheses are true, FDR is smaller or equal to FWER. Consequently, a method which controls FDR can be less stringent than a method controlling FWER [64]. The Benjamini-Hochberg correction method ranks the *P* values by ascending values and identifies the largest k such that the requirement in equation 2.21 is fulfilled. All null hypotheses with ranks below k are rejected [64].

$$P_k \leq \frac{k}{m} q^* \tag{2.21}$$

where m is the number of tested null hypotheses and q^* is the chosen FDR.

Chapter 3

Method

This chapter will present the material and methods used in this thesis. It will first describe the approaches used to develop and test methods for creating consensus CSD networks. Next, it will describe the application of the developed methods to create a consensus CSD network for BP. The approaches used to analyse and evaluate the resulting network(s) will also be presented. A third section is included to provide an overview of the relevant software for this thesis.

3.1 Method Development: Consensus Networks

One of the aims in this thesis is to develop and test methods for combining several data sets into one consensus CSD network. A potential method for creating such networks is to combine correlation coefficients from individual studies into averaged values, which subsequently are used as input to the conventional CSD approach. Here, two methods for estimating combined correlation coefficients in a network setting will be evaluated. This assessment will be based on tests performed on a large data set which, due to its large sample size, may be used as a relatively accurate reference set. Consequently, a gene expression data set named "*Skin - Not Sun Exposed (Suprapubic)*" (with 25 279 Gencode IDs and 517 samples) was downloaded on 25.08.21 from the Genotype-Tissue Expression (GTEx) consortium [65] (https://gtexportal.org). The chosen data set contained Gencode IDs as identifiers, which were translated to gene names using the translation file provided by the GTEx Portal (https://gtexportal.org/home/datasets; file: gencode.v26.GRCh38.genes.gtf). If two or more Gencode IDs corresponded to the same gene name, their expressions were averaged.

To reduce the running time of the analysis, and thus allow repeated testing, the number of genes in the data set was reduced to 1000. These were randomly chosen from the gene expression data set. The Spearman rank correlation coefficients were calculated for the 1000-gene data set as described by the CSD approach, see description of CSD. The subsampling was omitted and variances were not calculated. The resulting correlation coefficients will be used as references when testing the methods for estimating combined correlation coefficients (see details below).

The 1000-gene data set was subsequently divided into random subgroups, each containing between 10 and 49 samples. Note that these individual subgroups are deemed insufficient for CSD analysis due to their small sample sizes. In special cases, a maximum of 58 sam-

ples was allowed in order to include all samples from the entire 1000-gene data set in the subgroups. The random splitting process was repeated 100 times and Spearman rank correlations were calculated for all subgroups (without calculation of variances). Finally, two combined correlation coefficients were calculated for each gene pair within each repetition. These combined correlation coefficients were based on averages of Fisher's Z transformed correlation coefficients (equation 2.15, 2.16, 2.17) and weighted untransformed correlation coefficients (equation 2.18), respectively. Note that if a Spearman rank correlation coefficient is equal to 1, the Fisher's Z value will approach infinity. In these cases, Fisher's Z was set equal to 5 in our method (corresponding to a Spearman rank correlation coefficient of 0.9999). Similarly, a Spearman correlation coefficient of -1 was set equal to a Fisher's Z value of -5.

The estimated combined correlation coefficients, as well as individual subgroup correlation coefficients, were compared with the reference correlation coefficients calculated directly from the entire 1000-gene data set. Self-correlations were excluded from the comparisons. The following calculations and tests were performed:

1. Calculation of Spearman rank correlation between test (combined or subgroup) and reference correlation coefficients. The statistical difference between the correlation of correlations for the combination methods was evaluated by the Wilcoxon signed-rank test. Notice that the Spearman rank correlation coefficient is calculated even though it is expected a linear relationship between reference and estimated values. This is due to the inclusion of averaged values for the variables in the equation for the Pearson correlation coefficient (see equation 2.9).

2. Calculation of RMSE between test (combined or subgroup) and reference correlation coefficients. The statistical difference between the RMSEs for the combination methods was evaluated by the Wilcoxon signed-rank test.

3. Calculation of Jaccard indices as a function of number of investigated gene pairs between test (combined or subgroup) and the reference correlation coefficients. Absolute values of the correlations were used as input. The Jaccard indices between the two combination methods were also calculated in this step.

Together, these tests allow evaluation and comparison of the estimated combined correlation coefficients with each other and with individual subgroup correlation coefficients (representing the current method for dealing with small sample sizes in the CSD approach). A flowchart of the resulting method is shown in Figure 3.1.

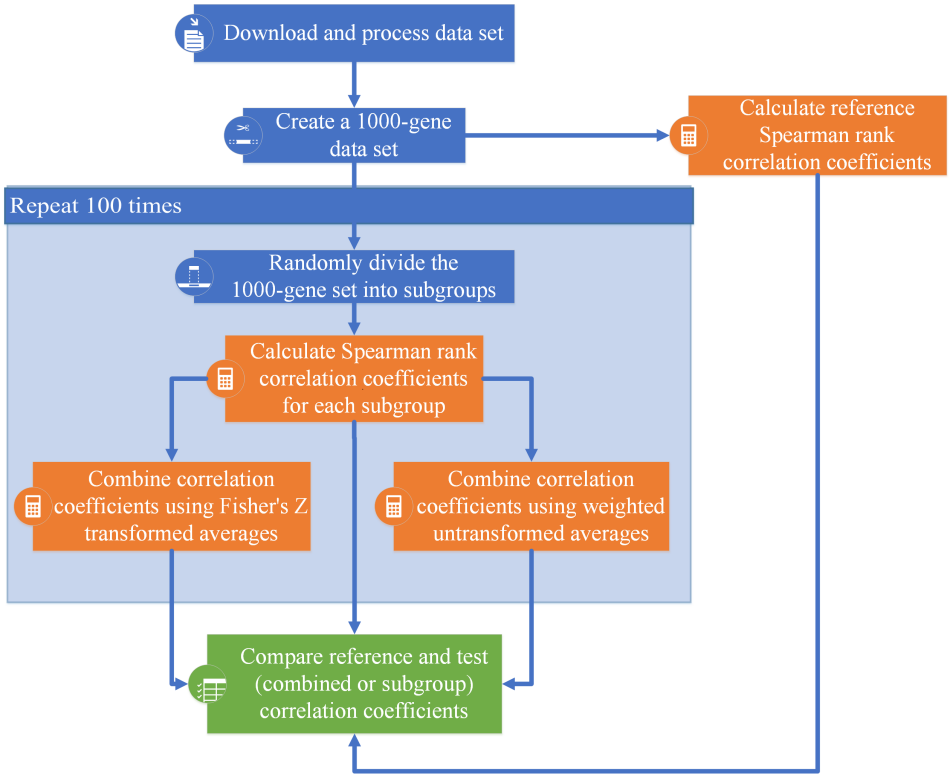


Figure 3.1: Flowchart of the method development for comparison of two methods for estimating combined correlation coefficient. Blue boxes indicate processing steps, orange boxes indicate calculation steps and the green box represents the comparison steps.

3.2 Network Analysis of Bipolar Disorder

In this section, the developed method from Section 3.1 is used to generate consensus CSD network(s) for BP. The necessary steps for the creation and analyses of the networks, as well as comparisons of the CSD networks based on the combination approaches, are presented in Figure 3.2. The following subsections will discuss each step in more detail.

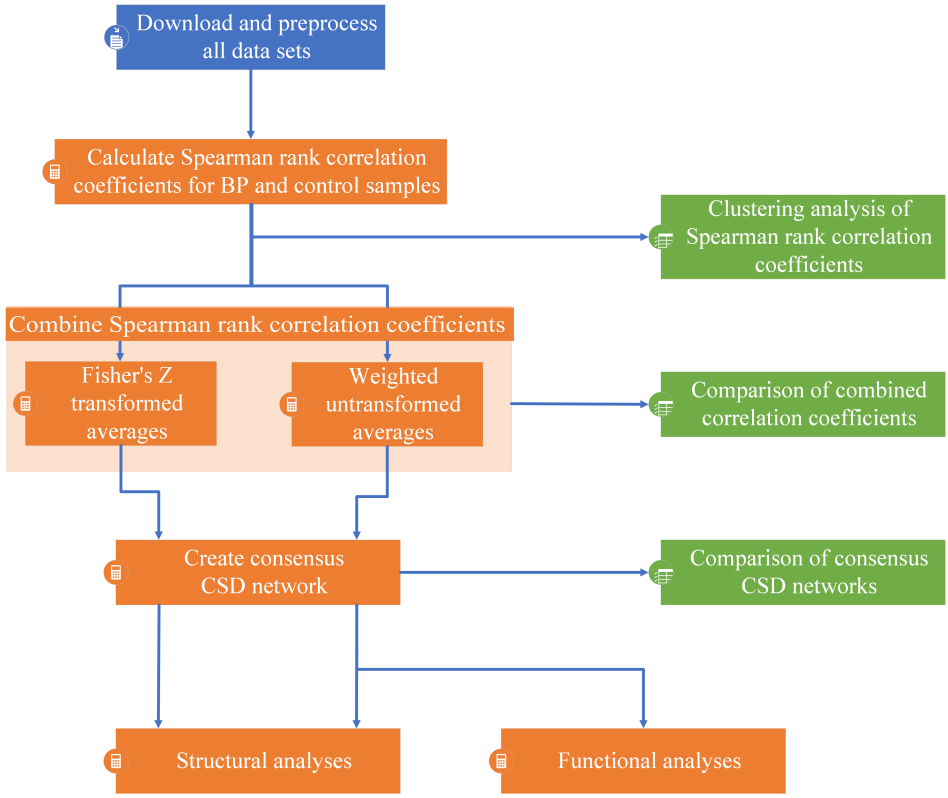


Figure 3.2: Flowchart of the method for creating consensus CSD networks for bipolar disorder (BP). The blue box indicates processing steps, orange boxes indicate calculation steps and green boxes represent the comparison steps. Note that the two parallel paths after the combination of Spearman rank correlation coefficients represent the creation and analyses of CSD networks based on either Fisher's Z transformed or weighted untransformed averages of correlation coefficients.

3.2.1 Network Construction

The first step of the consensus CSD network construction entails the identification of relevant BP data sets. Hence, a search for ”bipolar disorder” was performed in Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/gds) [66, 67] on 09.09.21. Results were restricted to ”Expression profiling by array” or ”Expression profiling by high throughput sequencing”. The search was further restricted to data sets with a total of at least 20 samples from humans. 45 search results matched these restrictions. A manual selection of data sets with at least 10 samples for both BP and control originating from macrodissected DLPFC was performed. The included studies and some of their characteristics are provided in Table 3.1. A summary of normalization methods for each study is provided in Appendix A.1, where only data set GSE80655 [68] required additional normalization. Note that none of the identified studies contained sufficient number of BP samples to be appropriate for CSD analysis alone.

Each relevant data set was downloaded and split into two groups, containing samples from either BP or controls. If the studies included samples from other tissues in addition to the DLPFC or samples from other disorders, these additional samples were omitted. The probe IDs were converted to gene names using the provided files at GEO or an appropriate translation file from BioMart (https://www.ensembl.org/biomart, downloaded 29.10.21). Consequently, a common naming regime for the included studies was created. Probes that corresponded to more than one gene were removed from the data sets. If two or more probes corresponded to the same gene, these were also omitted. This is due to the unknown consequence of combining averaged expressions from one data set with possible unaveraged expressions from another data set when estimating combined correlation coefficients. Furthermore, genes with expression values that did not meet the requirements for calculation of Spearman rank correlation coefficients were also removed. Specifically, this corresponds to genes where all measured expression values were equal to zero. Thus,

Table 3.1: Included studies of bipolar disorder from Gene Expression Omnibus (GEO). All studies contained samples originating from the dorsolateral prefrontal cortex.

GEO accession	Number of BP samples	Number of control samples	Provider of brain tissue	Reference
GSE80655	23	24	Pritzker Neuropsychiatric Disorders Research Consortium	[68]
GSE92538	26	83	Pritzker Neuropsychiatric Disorders Research Consortium	[69]
GSE53987	17	19	University of Pittsburgh	[70]
GSE5388	30	31	Stanley Medical Research Institute	[71]
GSE12649	33	34	Stanley Medical Research Institute	[72]
GSE120340	10	10	Stanley Medical Research Institute	[73]
Total	139	201		

they could not be ranked. This finding was only apparent in the GSE80655 data set [68]. Finally, genes that were not universally represented in all six data sets removed. This gave a total of 3148 genes that were analysed in the subsequent steps. The resulting expression data sets were sorted to make sure that the genes were in the same order. The Spearman rank correlation coefficients were calculated for all data sets individually as described in the conventional CSD approach, see description of CSD. However, the variances were not calculated due to low number of samples in each individual study.

Table 3.1 indicates that the brain tissues in some studies originated from the same brain bank. A mapping of available patient information (Appendix A.2) from these studies suggests that they might rely on (some of) the same patients. Hence, a clustering analysis was conducted to investigate the similarity between the data sets. This analysis relied on the creation of matrices where each element corresponded to a pairwise Spearman rank correlation coefficients between the Spearman rank correlation coefficients from two studies. The matrices for BP and control samples were used to generate hierarchically-clustered heat maps for the data sets.

The Spearman rank correlation coefficients from each individual study were subsequently combined using the methods from Section 3.1 (combined correlation coefficients based on Fisher's Z transformed or weighted untransformed averages). These results were further processed to fit the required format for subsequent CSD analysis and led to the production of four data sets:

1. Combined correlation coefficients for BP samples based on Fisher's Z transformed values.
2. Combined correlation coefficients for control samples based on Fisher's Z transformed values.
3. Combined correlation coefficients for BP samples based on weighted untransformed values.
4. Combined correlation coefficients for control samples based on weighted untransformed values.

In the last steps, the CSD scores between BP and control samples were calculated and a final CSD network was generated using an importance level of $p = 10^{-4}$ (selSize = 10 000). These two last steps were performed twice, using combined correlation coefficient estimates based on either Fisher's Z transformed averages or weighted untransformed averages.

As the p in the CSD approach does not correspond to a significance threshold, it was wishful to evaluate the significance of the included links in the CSD network. Each link is associated with two correlation coefficients, one originating from BP samples and one from control samples. In each case, the null hypothesis that there was no correlation in the population was tested using a two-tailed t-test with degrees of freedom (df) $= n - 2$, as recommended in [34]. Next, the *P* value for each link was calculated as the product of the *P* values for its two underlying correlation coefficients. As several hypotheses have

been tested, it was necessary to correct for multiple comparisons. Hence, the Benjamini-Hochberg correction (with FDR < 0.05) was applied. Links that did not fulfill the significance requirement were removed from the CSD network.

3.2.2 Comparison of Models

The method development in Section 3.1 was based on combination of gene expressions which originated from the same underlying data set. In contrast, the network analysis of BP has been based on gene expressions from different underlying data sets. Consequently, a comparison of the estimated correlation coefficients was also conducted for these data sets. This was performed by calculating the Spearman rank correlation coefficient for estimates of correlation coefficients based on Fisher's Z transformed averages relative to combined scores based on weighted untransformed averages. Obvious outliers (where the difference was greater than 0.4) were investigated manually. The combination methods were further compared by calculating their Jaccard indices for different number of investigated gene pairs. Jaccard indices were also calculated between individual BP and control data sets relative to the combined data sets.

As a follow-up, the Jaccard indices between nodes and links in the the two CSD networks (based on either Fisher's Z transformed or weighted untransformed values) were calculated. Similar calculations were also conducted for each of the network subtypes from the CSD analysis (C, S and D networks). Furthermore, the CSD networks were imported into Cytoscape [74] for visualization. A merged network of the CSD networks was created and used to visualize commonalities and differences between the networks based on Fisher's Z transformed and weighted untransformed averages. The merged network was also visualized by plotting its adjacency matrix, where the nodes were first sorted by network origin and then by community membership determined by the Louvain algorithm [29]. The CSD networks were further compared node by node to investigate the concordance between the specific degrees as well as the identities of the neighbours. The similarity between the neighbours of one node in the two CSD networks was measured by calculating the Jaccard index.

3.2.3 Network Analysis

The final CSD networks were subjected to structural network analyses to investigate if they behave as expected from [9]. The structural analyses treated the networks as undirected and unweighted. These analyses included investigation of the degree distributions for the CSD networks, as well as calculation of network assortativity and average clustering coefficients for the CSD networks and the individual C, S and D networks. The calculated assortativity and average clustering coefficients were compared with values from random networks generated according to the configuration model [27]. Consequently, it was investigated if the the calculated assortativity and average clustering coefficients were significantly different from random expectations. The estimated P values were corrected for multiple testing with the Bonferroni correction method.

In the next step, homogeneity scores were calculated for all nodes in the networks to assess if nodes tended to be connected with one or several link types. Genes were subsequently classified according to degree, where $3 \leq k \leq 9$ are defined as intermediate genes and $k \geq 10$ as hubs. The tendency of hubs to be more homogeneous than intermediate genes was tested with a one-tailed independent samples t-test. In addition, the Spearman rank correlation coefficient between homogeneity and degree was calculated for all genes with $k \geq 3$.

In addition to structural analyses of the networks, functional analyses were also conducted. These analyses were restricted to the CSD network based weighted untransformed averages of correlation coefficients. All nodes of the individual C, S and D networks, as well as the full CSD network, were subjected to GO enrichment analyses using the analysis tool from PANTHER [61, 75, 76]. Gene names of the nodes were used as input to the GO enrichment analysis. If a gene name was not uniquely mapped, a manual inspection of possible matches was performed and the correct ID was chosen. The analyses were restricted to the complete GO biological process using the list of all included genes in the CSD analysis as reference. The test type was set to Fisher's exact test with FDR correction.

The CSD network was also subjected to a disease enrichment analysis in Cytoscape [74], using the DisGeNET Cytoscape app [77, 78] with gene symbols as input. DisGeNET integrates data from several databases to provide information about the current knowledge related to the genetic basis of human diseases [77]. The DisGeNET enrichment analysis uses Fisher's tests and corrects *P* values for multiple testing by the Benjamini-Hochberg method [78]. However, the disease enrichment analysis for the CSD network has been restricted to overrepresentation of genes related to BP. Hence, no correction for multiple testing was required. The enrichment analysis was further restricted to curated information.

In the next analysis step, communities in the CSD network were investigated. The communities were identified using the Louvain algorithm [29]. Communities that contained more than 5 nodes were subsequently subjected to GO enrichment analysis, using the same settings as described above.

Finally, the top hubs in the CSD network were investigated in more detail. This included a manual investigation of relevant genes and underlying correlations between these genes and their neighbours. The purpose of the latter was to identify the direction of change for correlations underlying D links and whether BP samples gained or lost a correlation in the case of S links. This inspection was subsequently extended to all S links.

3.2.4 Comparison with Basal Ganglia

The correlations underlying the final CSD network (based on weighted untransformed averages) were compared with correlations from basal ganglia. The idea was to examine if the co-expressions of genes in the DLPFCs from BP samples were shifted towards co-expression patterns in another brain region relative to the control DLPFCs. The basal ganglia was chosen due to the availability of CSD data from this brain region in [9] (num-

ber of samples: 92, available from Almaas Lab as CSD_complete_data.txt.gz).

For each gene pair, a binary test was conducted to check if the sign of the correlation difference between control DLPFC and BP DLPFC was equal to the sign of the difference between control DLPFC and basal ganglia. No transformation was required at this step as only the signs of the differences were of interest. The number of gene pairs that fulfilled the binary test was counted for all links in the CSD network that had a counterpart in the basal ganglia data set. The links were also sorted according to their link types to allow more detailed comparisons. The *P* values for all comparisons were calculated by comparing the number of successes to the hypothesized probability of success $(2/3)$ using a two-tailed binomial test. As four hypotheses have been tested, the calculated *P* values were adjusted for multiple testing with the Bonferroni correction method.

3.3 Software

The development, testing, network creation and analysis in this thesis were mainly performed in Python. Plots were created using matplotlib [79] and Seaborn [80], and venn diagrams were generated using matplotlib-venn (matplotlib_venn). Processing, some data analysis and scientific computations relied on Pandas [81] and Numpy [82]. Performance evaluations and network analyses were conducted using SciPy [83] and NetworkX [84]. In addition to computations and analyses in Python, C++ and DESeq [85, 86] in R were used to calculate gene pair correlations and to normalize the data set GSE80655, respectively. Furthermore, Excel was used for significance filtering of the final CSD networks.

The CSD networks were visualized and subjected to simple network analysis using Cytoscape [74]. The Cytoscape-plugins DisGeNet-app [77, 78] and setsApp [87] have been used to perform disease enrichment analysis and to aid the community layout of the CSD network, respectively. GO enrichment analyses have been conducted in PANTHER [61, 75, 76]. All relevant code, as well as the CSD networks, are available on GitHub (CSD and Consensus_CSD). An overview of version numbers of the software is provided in Appendix B. In addition, all processed data sets are provided at Figshare (DOI: https://doi.org/10.6084/m9.figshare.19665624.v1).

Chapter 4

Results and Analysis

This chapter will present the results of this thesis and an initial interpretation and analysis of the findings. The first section investigates and compares the methods designed to construct CNs from combined correlations. The second section presents the consensus CSD network(s) for BP and includes structural and functional analyses of the network(s).

4.1 Method Development: Consensus Networks

The current method for dealing with small sample sizes in the CSD approach is to carry out the analysis with just one data set and omitting the calculation of variances. One of the aims in this thesis is to develop and test new methods for improving this approach to allow creation of consensus CSD networks. Two strategies for estimating combined Spearman rank correlation coefficients from several underlying subgroups, based on either Fisher's Z transformed or weighted untransformed averages of correlation coefficients, have been evaluated for this purpose. The following subsections will describe the comparison of these estimates and individual subgroup correlations to reference Spearman rank correlation coefficients from the 1000-gene set originating from the data set called "*Skin - Not Sun Exposed (Suprapubic)*".

4.1.1 Correlation of Correlations

Spearman rank correlation coefficients from the combination methods and individual subgroups were compared to reference Spearman rank correlation coefficients by calculating their relative Spearman rank correlation coefficients. This is a quite intricate description and will be simplified by referring to this metric as "correlation of correlations". This comparison allows evaluation of changes in ranking of gene pairs across the entire gene set.

Figure 4.1 shows combined and subgroup correlation coefficients plotted against the reference correlations for representative repetitions and subgroups (see Figure 3.1 for a reminder that there was carried out 100 repetitions, each with several subgroups). Representative plots were selected based on closeness to the median correlation of correlations within each group. Figure 4.1 illustrates that the concordance between the subgroups and reference set increases as the sample size of the subgroups increases. Nevertheless, both estimation methods for combined correlations clearly outperform the subgroups when compared to the reference set. This is seen in Figure 4.1e and 4.1f, where the plotted heat

maps approach straight lines and have high correlations of correlations (0.9961 for both combination methods). The general trends from Figure 4.1 were reproduced in all repetitions. This is shown in Figure 4.2 where the test method (subgroups of different sizes and combination methods) is plotted against its correlation of correlations. Once again, it is clear that the combination methods outperform the individual subgroups as estimators of the reference correlations.

At first glance at Figure 4.1e, 4.1f and 4.2b, it does not appear to be an obvious difference between the two combination methods for estimating the reference correlations. As the estimates can be paired for each repetition, the pairwise differences have also been evaluated to generate a more detailed comparison (Appendix C.1). In general, the correlation of correlations tends to be higher for the weighted untransformed averages than the Fisher's Z transformed averages, but the difference is only apparent in the fifth decimal place. Hence, the difference between the estimation methods is assumed to have minor impact and is given little importance when comparing the combination methods despite being significant (Wilcoxon signed-rank test: $P = 3.9 * 10^{-5}$).

In summary, the estimation of combined correlation coefficients seems to outperform the current approach for dealing with small sample sizes in the CSD analysis when evaluating correlation of correlations. Weighted untransformed averages of correlation coefficients have significantly higher correlation of correlations than Fisher's Z transformed averages when comparing pairwise estimates, but the difference is small.

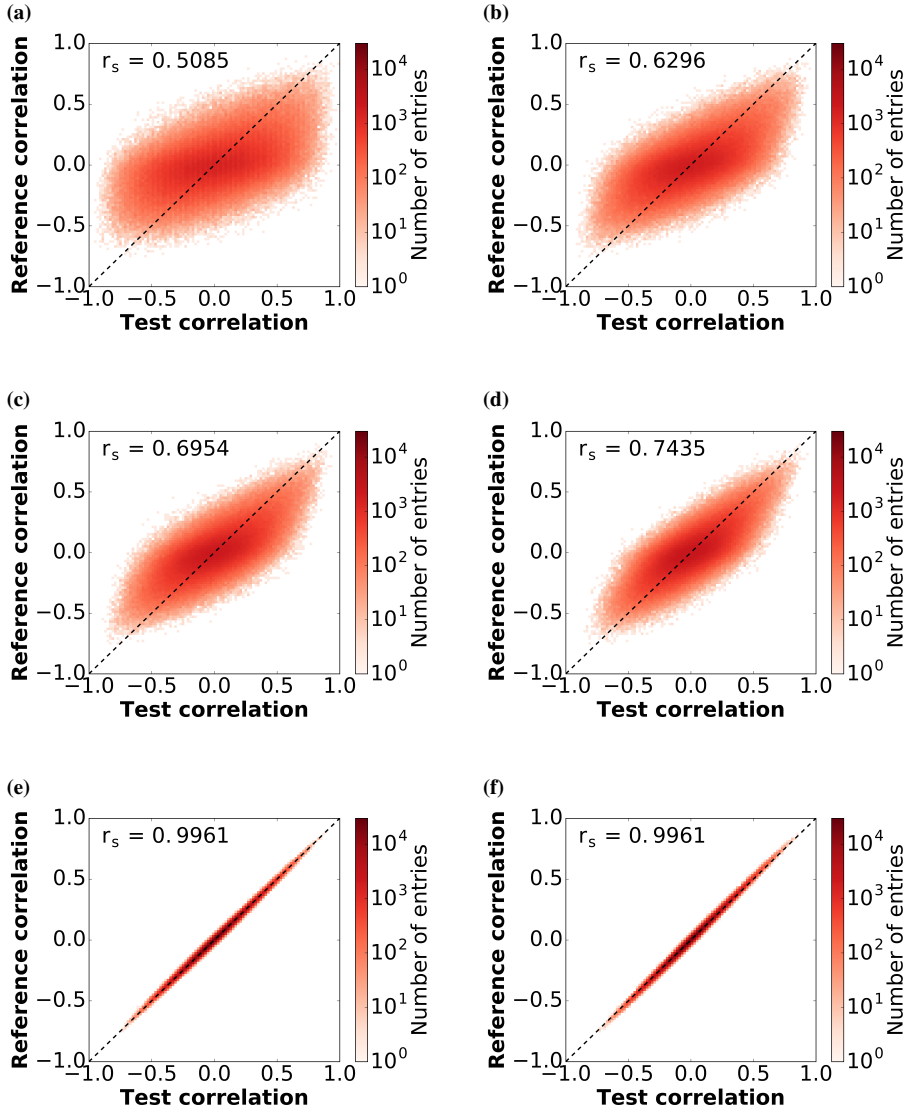


Figure 4.1: Representative heat maps between test and reference Spearman rank correlation coefficients. (a)-(d) are representative heat maps where the test data sets correspond to subgroups of sizes 10-19, 20-29, 30-39 and 40-49, respectively. (e) and (f) are representative heat maps where the test data sets correspond to combined correlation coefficients based on Fisher's Z transformed and weighted untransformed averages, respectively. The dashed lines correspond to the expected relationship ($y = x$). r_s represents the Spearman rank correlation coefficient between the test and reference correlations.

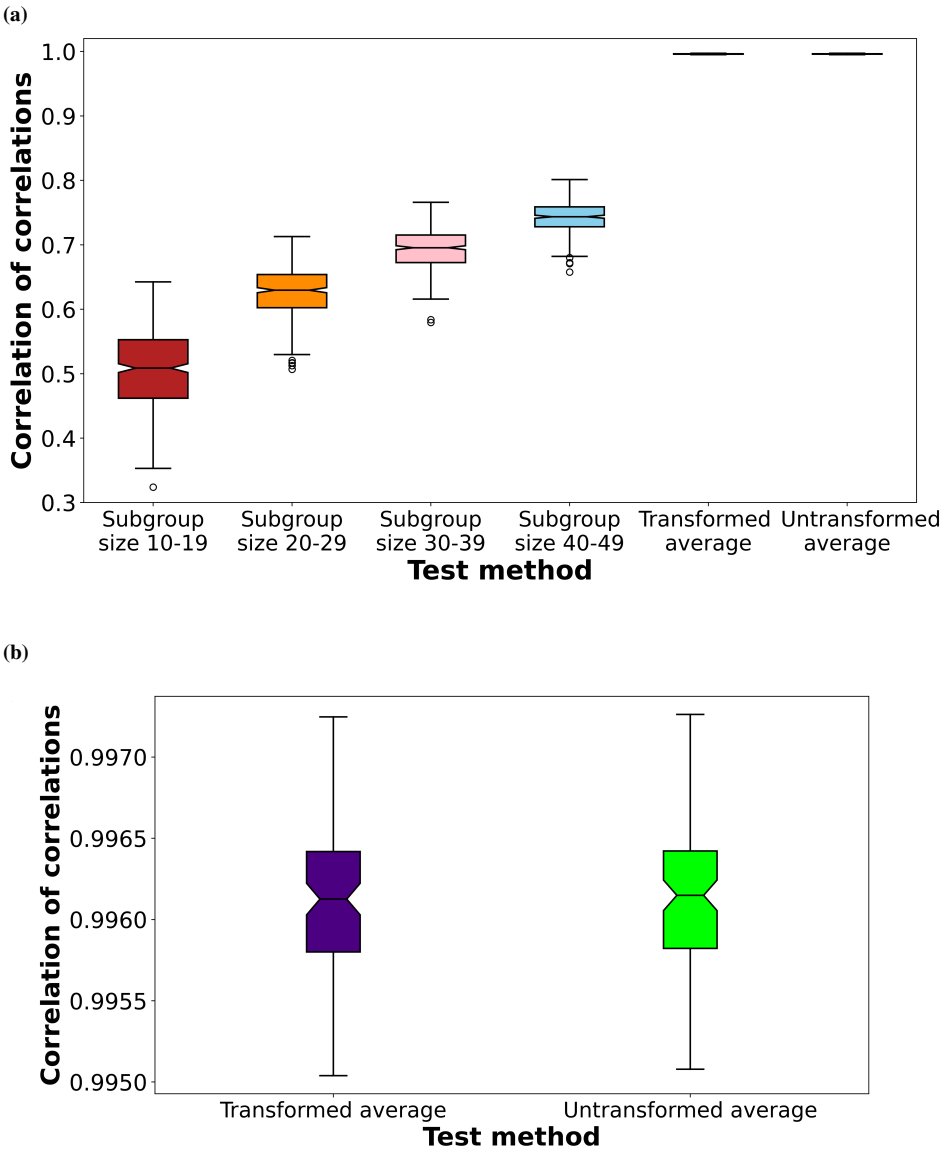


Figure 4.2: Box plots of Spearman rank correlation coefficients between test and reference correlation coefficients, where (a) test correlations originate from subgroups of indicated sizes or combined values using Fisher's Z transformed averages or weighted untransformed averages. (b) Comparison of Spearman rank correlation coefficients between combined and reference correlation coefficients at a finer scale.

4.1.2 Root Mean Square Error

The second comparison of combined and subgroups Spearman rank correlation coefficients entailed the calculation of their RMSEs relative to the reference Spearman rank correlation coefficients. A box plot of the calculated RMSEs is presented in Figure 4.3. As expected, the RMSE tends to decrease as the subgroup size increases. This indicates a better fit between the Spearman rank correlation coefficients of the subgroups and the reference data set as the sample size increases. Interestingly, both combination methods clearly outperform the subgroups and display substantially lower RMSEs.

Figure 4.3b indicates a small difference in RMSEs between the combination methods. A more detailed comparison of the pairwise differences between these methods is provided in Appendix C.2. It appears as the Fisher's Z transformed averages tend to generate lower RMSEs than the weighted untransformed averages. However, these differences are only apparent in the third decimal place and is therefore believed to have minor importance for the choice of estimation method, despite being significant (Wilcoxon signed-rank test: $P = 4.0 * 10^{-18}$).

In total, the calculated RMSEs between test and reference Spearman rank correlation coefficients support the use of estimated combined correlation coefficients as an alternative approach to the current method for dealing with small sample sizes in CSD analysis. The calculated RMSEs indicate a better fit between the reference Spearman rank correlation coefficients and Fisher's Z transformed averages than weighted untransformed averages, but the difference is small.

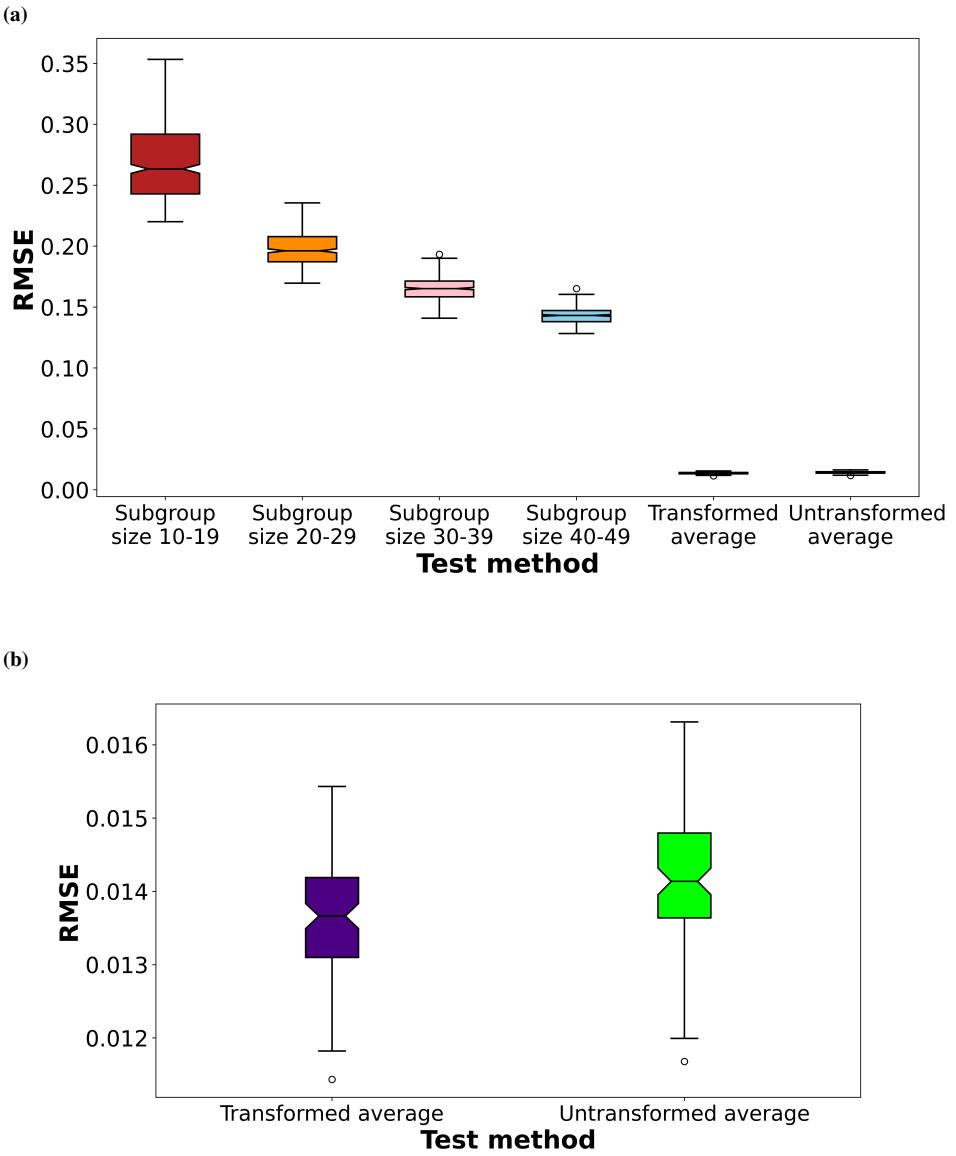


Figure 4.3: Box plots of root mean square error (RMSE) between test and reference correlation coefficients. (a) Comparison of RMSEs where test data sets correspond to subgroups of varying sizes or combined data sets (based on Fisher's Z transformed or weighted untransformed averages of correlation coefficients). (b) Comparison of RMSEs at a finer scale where test data sets correspond to combined data sets.

4.1.3 Jaccard Index

The third and final comparison of the combined and subgroup correlation coefficients involved the calculation of Jaccard indices for top n gene pairs with highest absolute value of correlation coefficients relative to the reference data set. This comparison allows evaluation of changes in ranking of the most strongly co-expressed gene pairs, which are the foundation of CSD analysis.

The Jaccard index between test and reference Spearman rank correlation coefficients is plotted as a function of number of investigated gene pairs in Figure 4.4. In general, increasing sample size tends to increase the Jaccard index. Furthermore, the Jaccard indices are higher for combined correlation coefficients than individual subgroups for all numbers of investigated gene pairs. From a visual inspection of the plot, there is no obvious difference between the combination methods based on Fisher's Z transformed and weighted untransformed averages. A plot of Jaccard indices calculated between top n gene pairs of the Fisher's Z transformed averages relative to the weighted untransformed averages is included in Appendix C.3. This plot supports the notion that there is a high degree of similarity between the combination methods.

Interestingly, there is a drop in Figure 4.4 for both combination methods when the number of investigated gene pairs is 10. This could simply be caused by stochastic effects. When the number of investigated gene pairs increases above 30, the Jaccard indices stabilize and remain above 0.8 for both combination methods. As a CSD network typically includes around 1000 gene pairs [9], the drop and large variation of the Jaccard indices at low number of investigated gene pairs are expected to have minor influence on the final result. In addition, the combination methods still perform better than the individual subgroups even at low number of investigated gene pairs.

The analysis of Jaccard indices supports the observations from the investigation of correlation of correlations and RMSEs – The use of combined correlation coefficients is a superior method compared to the use of individual subgroups for estimating reference correlations. The analysis of correlation of correlations weakly favours weighted untransformed averages, while the RMSE analysis weakly favours Fisher's Z transformed averages. The Jaccard index analysis does not favour any of the combination methods above the other. Consequently, both Fisher's Z transformed and weighted untransformed averages seem as viable estimation methods for combined correlation coefficients.

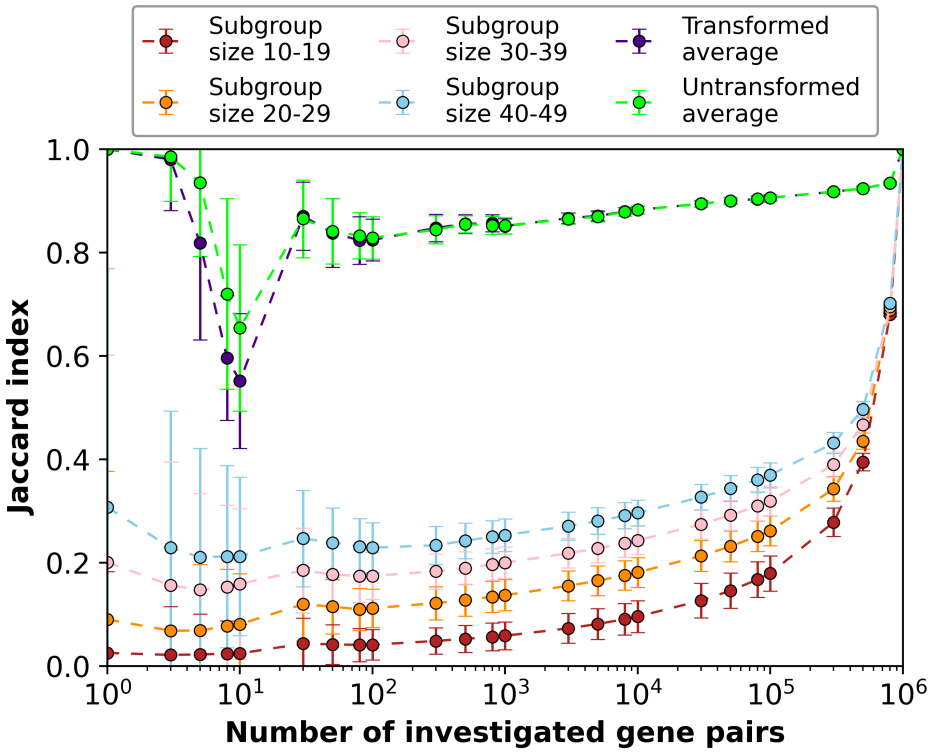


Figure 4.4: Jaccard index as a function of number of investigated gene pairs. Different methods for estimating test correlation coefficients have been compared to the reference Spearman rank correlation coefficients. Transformed and untransformed averages refer to Fisher's Z transformed and weighted untransformed averages, respectively. Error bars represent standard deviations.

4.2 Network Analysis of Bipolar Disorder

The following section is devoted to the network analysis of BP. First, the results of the clustering analysis of the data sets will be presented. Furthermore, as the method development relied on data from just one data set, this section will also compare combined correlation coefficients based on Fisher's Z transformed and weighted untransformed averages for BP and control samples. Subsequently, the focus will shift to comparison and structural analysis of the generated CSD networks (based on either Fisher's Z transformed or weighted untransformed averages). In the final parts of this section, functional analyses of the CSD network(s), its communities and hubs will be presented. This will also include a presentation of functions and potential disease contributions of central genes.

4.2.1 Cluster Analysis of Data Sets

The network analysis of BP has relied on the combination of six individual data sets. As indicated in Table 3.1, two studies (GSE80655 [68], GSE92538 [69]) have received brain

tissue from Pritzker Neuropsychiatric Disorders Research and three studies (GSE5388 [71], GSE12649 [72], GSE120340 [73]) have received brain tissue from Stanley Medical Research Institute. Comparisons of the available information from these studies are provided in Appendix A.2. GSE80655 [68] and GSE92538 [69] have a relatively good match between age, gender and ethnicity of their included patients. Similarly, there is a quite striking similarity between age, gender and age of onset of BP for patients from GSE5388 [71] and GSE12649 [72] (no available information from GSE120340 [73]). This could suggest that studies that receive tissue from the same brain bank are based on (some of) the same patients. Hence, a clustering analysis was conducted to investigate the similarity between the data sets.

The clustering analysis of the data sets underlying BP and control samples relied on comparisons of Spearman rank correlation coefficients from the individual studies by calculating their pairwise correlation of correlations. The resulting hierarchically-clustered heat map for the BP data sets is shown in Figure 4.5. A similar plot for control samples is included in Appendix D. The plots do not indicate a clear clustering of data sets that originate from the same brain bank. Hence, data sets originating from the same brain bank do not appear to be more similar to each other than the other data sets. Consequently, all data sets will be treated as independent in this thesis.

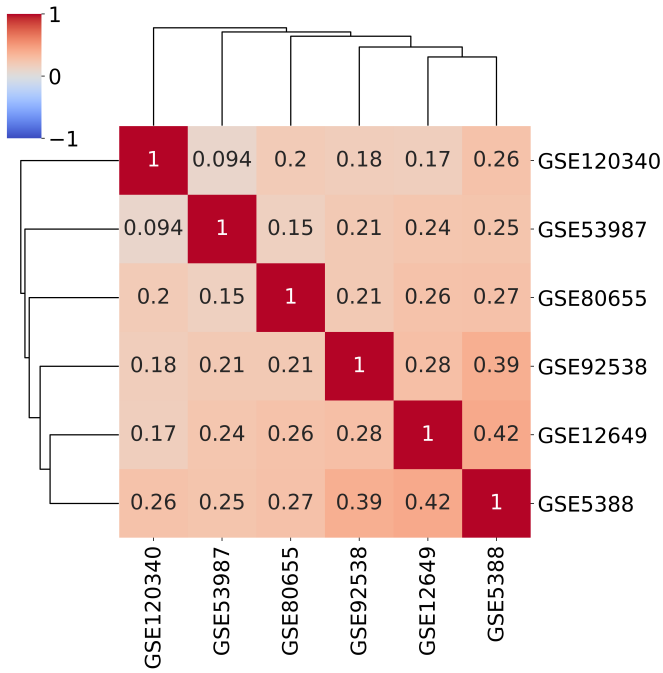


Figure 4.5: Hierarchically-clustered heat map for bipolar disorder data sets. The elements of the matrix correspond to the pairwise Spearman rank correlation coefficients between the Spearman rank correlation coefficients from the two indicated studies.

4.2.2 Model Comparison at the Level of Correlations

The construction of a consensus CSD network for BP included an intermediate step where Spearman rank correlation coefficients between gene expressions in each conditions were calculated and average values were estimated. This subsection presents the results from the comparison of the two methods for estimating these averages. A total of 3148 genes were included in these analyses.

The methods for estimating combined Spearman rank correlation coefficients from individual BP data sets are plotted against each other in Figure 4.6. A similar plot for the control samples is given in Appendix E.1. The Spearman rank correlation coefficients between the combination methods are 0.9991 and 0.9993 for BP and control samples, respectively. Together, this illustrates that there generally is a good agreement between estimates based on Fisher's Z transformed and weighted untransformed averages of correlation coefficients.

A visual inspection of Figure 4.6 shows that the plot has a weak S-shape. This indicates that the Fisher's Z transformed averages typically generate more extreme estimates than the weighted untransformed averages. Hence, the latter can be considered a more conservative method. A manual inspection of the outliers in Figure 4.6 (where the $|\text{difference}| > 0.4$) revealed that the underlying correlations of all the outliers included one perfect correlation. The perfect correlation hijacks the Fisher's Z transformed average, resulting in a larger estimate compared to the weighted untransformed average.

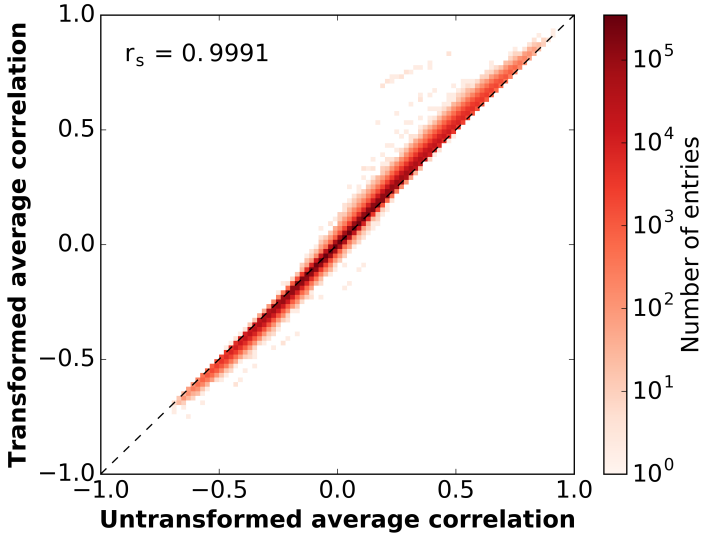


Figure 4.6: Heat map between combined Spearman rank correlation coefficients based on Fisher's Z transformed and weighted untransformed averages for bipolar disorder. The dashed line represents the expected relationship ($y = x$). r_s represents the Spearman rank correlation coefficient between the combined correlations.

The identified perfect correlations only appear between low-count genes in GSE80655 [68], where most of the samples have values equal to zero. Hence, the perfect correlations are not believed to represent important biological features. Consequently, the Fisher's Z transformed averages overestimate the importance of these gene pairs. As a result, weighted untransformed averages appear to be more suitable for CSD analysis than Fisher's Z transformed averages.

The top n gene pairs with highest absolute value of correlation coefficients (excluding self-correlations) from the combination methods were compared to each other and to the sample data sets by calculating Jaccard indices. This is illustrated in Figure 4.7 using combined correlation coefficients based on weighted untransformed values as reference. A similar plot using combined correlation coefficients based on Fisher's Z transformed values as reference is included in Appendix E.2. In general, the Jaccard indices are higher for weighted untransformed averages relative to Fisher's Z transformed averages than for combined correlation coefficients relative to sample correlations. This indicates that the combined correlation coefficients based on weighted untransformed or Fisher's Z transformed values are more similar to each other than to the individual underlying data sets. When comparing combined correlation coefficients to the BP sample correlations, the Jaccard indices between combined correlation coefficients relative to GSE5388 [71] and GSE12649 [72] have the highest values. This is expected as GSE5388 and GSE12649 contain the largest number of BP samples among the data sets. Similarly, the Jaccard indices between combined correlation coefficients relative to GSE92538 [69] have the highest values when comparing combined correlation coefficients to the control data sets. This is due to the large number of control samples in GSE92538.

Some general trends between weighted untransformed and Fisher's Z transformed averages may be observed in Figure 4.7. First, there is typically a lower Jaccard index between these combination methods when the number of investigated gene pairs is low. As noted in the results of the method development (Section 4.1.3), this may simply be caused by stochastic effects. As the number of investigated gene pairs increases, the Jaccard index also increases. When the number of investigated gene pairs is above 300, the Jaccard indices between the combination methods are always above 0.67 and 0.80 for BP and control samples, respectively. This indicates a substantial overlap between the top n gene pairs when n is above 300. The difference in Jaccard indices between BP and control samples may be caused by a different number of total samples or by differences in variation between the included samples.

Taken together, the Fisher's Z transformed and weighted untransformed averages of Spearman rank correlation coefficients between gene expressions from both BP and control samples generally exhibit good agreement both in terms of correlation of combined correlations and Jaccard indices. Interestingly, Fisher's Z transformed averages tend to generate more extreme estimates than weighted untransformed averaged and may be hijacked by spurious perfect correlations. Consequently, weighted untransformed averages seem to be a more suitable combination method for CSD analysis.

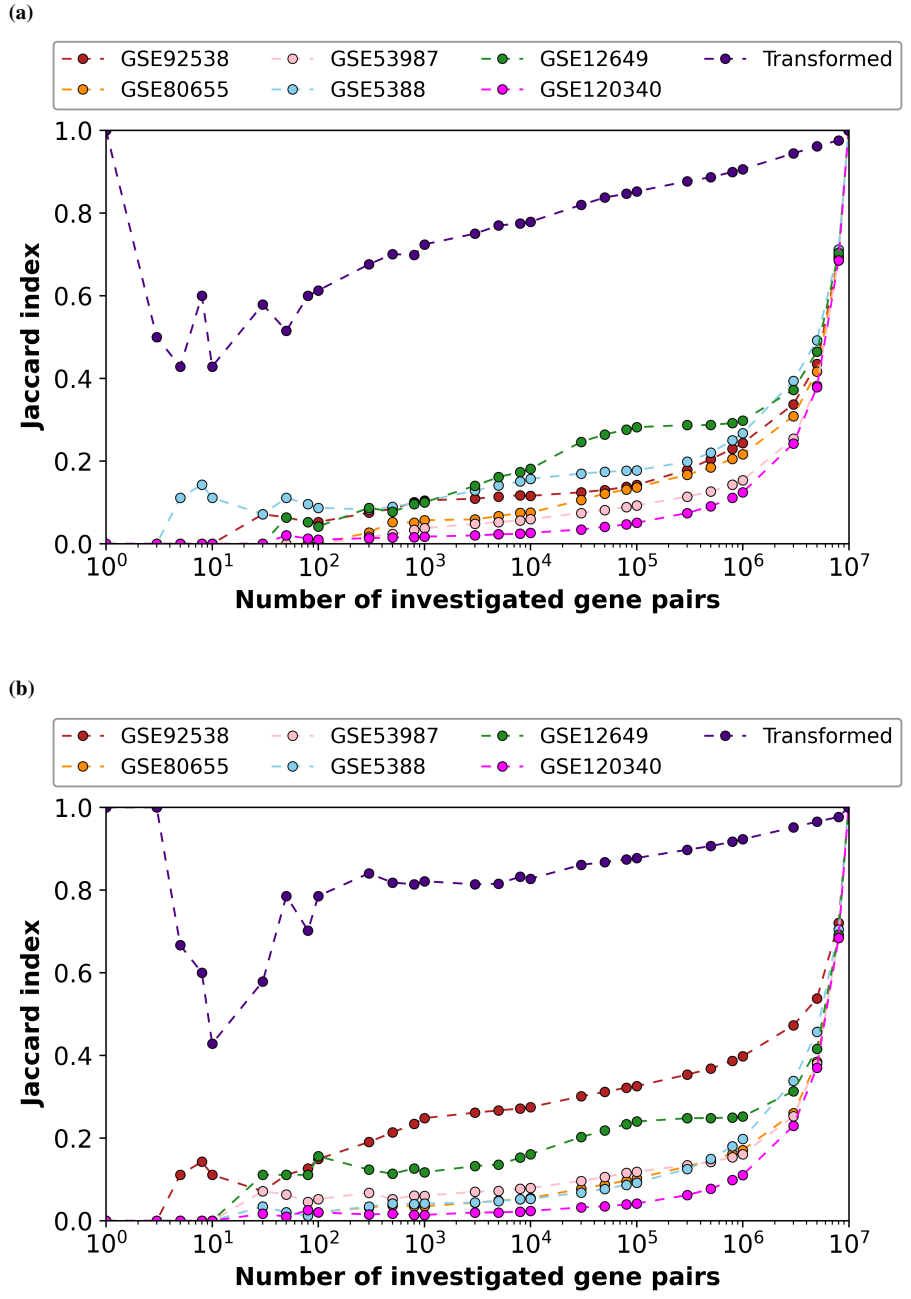


Figure 4.7: Jaccard index as a function of number of investigated gene pairs between Spearman rank correlation coefficients from indicated data sets and weighted untransformed averages originating from (a) bipolar disorder and (b) control samples. The term ”Transformed” refers to Fisher’s Z transformed averages of the correlation coefficients.

4.2.3 Model Comparison at the Network Level

The last steps of the CSD approach produce a final CSD network, with links defined as either C, S or D. In this thesis, two CSD networks have been created for BP. One of the networks is based on Fisher's Z transformed averages of correlation coefficients, while the other is based on weighted untransformed averages. In this subsection, the overlap and similarity between these final consensus networks (with $p = 10^{-4}$, followed by significance filtering) are evaluated.

The complete CSD network based on weighted untransformed averages contained 566 nodes and 747 links, while the network based on Fisher's Z transformed values contained 623 nodes and 811 links. These networks had 472 common nodes and 591 common links, where all common links were given the same link type (C, S or D) in the CSD networks. The calculated Jaccard indices were 0.66 and 0.61 for nodes and links, respectively. Jaccard indices for network subtypes are given in Table 4.1. All of these indices were calculated between the given network type based on weighted untransformed averages of correlation coefficients relative to the same network type based on Fisher's Z transformed averages. As seen in Table 4.1, the C networks have the highest Jaccard indices for both nodes and links. The S and D networks have comparable Jaccard indices for links, but the S networks have a slightly higher Jaccard index for nodes than the D networks.

A visualization of commonalities and differences between the CSD networks for BP is provided in Figure 4.8 and 4.9. As expected from Table 4.1, C links and nodes connected by C links are mostly shared between the CSD networks. Many S and D links are also shared, but to a lesser extent than the C links. Interestingly, the visualization indicates that nodes that are only present in one of the CSD networks are located in the periphery and have low degrees. Most of these nodes are not connected to the giant component (Figure 4.9). High-degree nodes on the other hand, are always defined as common between the networks. This interpretation is supported by Figure 4.10a, which illustrates that there is a good agreement between node degrees in the two CSD networks. It is noteworthy that the maximum degree of unique nodes, meaning nodes that are present in just one of the CSD networks, is four. This confirms the observation that unique nodes are located in the periphery of the merged network.

Table 4.1: Jaccard indices between nodes and links in the consensus CSD networks for bipolar disorder based on Fisher's Z transformed and weighted untransformed averages of correlation coefficients.

Network type	Jaccard index - nodes	Jaccard index - links
Full CSD network	0.66	0.61
C network	0.86	0.81
S network	0.61	0.50
D network	0.56	0.50

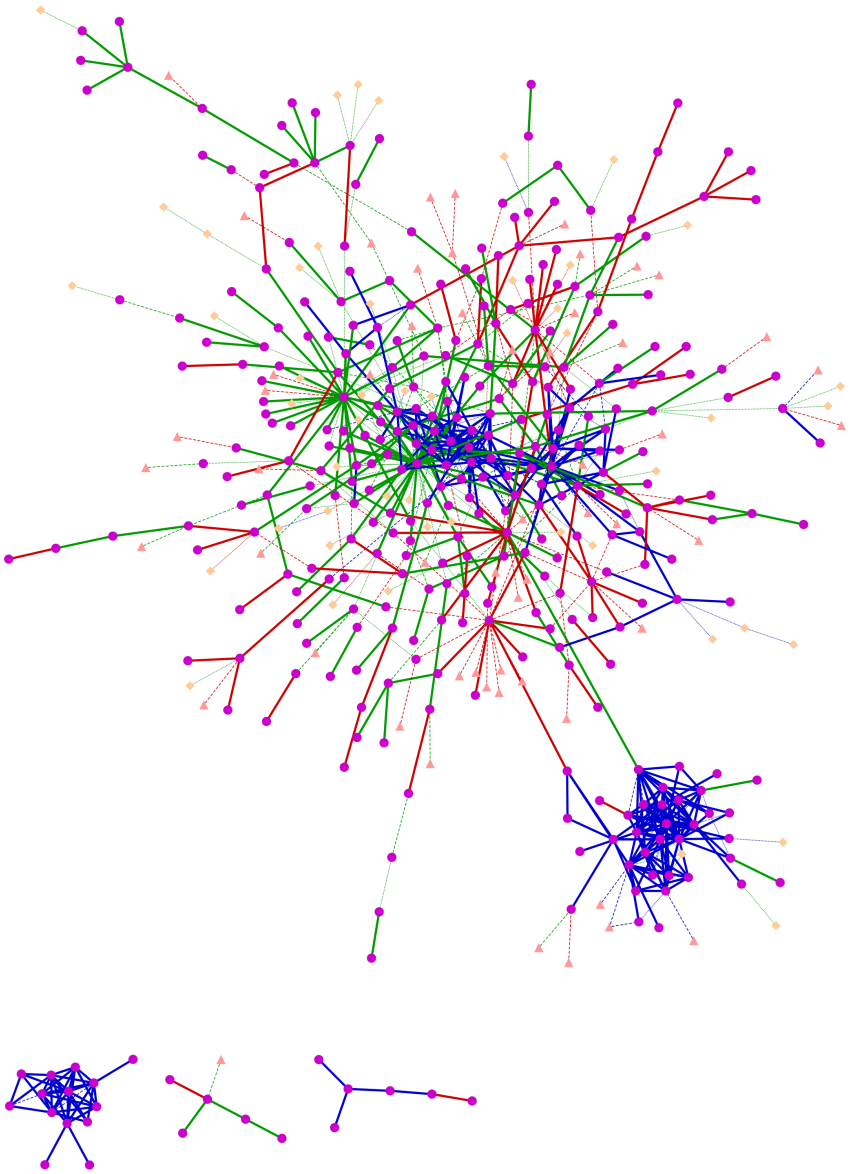


Figure 4.8: Graphical comparison of CSD networks for bipolar disorder based on Fisher's Z transformed and weighted untransformed averages of correlation coefficients. Components with five or fewer nodes have been excluded for illustration purposes. Nodes have been coloured and shaped according to their network origin where nodes unique for weighted untransformed averages, Fisher's Z transformed averages and common nodes are light orange diamonds, pink triangles and purple circles, respectively. Links have also been marked according to their network origin, where links unique for weighted untransformed averages, Fisher's Z transformed averages and common links are thin dotted, thin dashed and thick solid lines, respectively. The link colours represent the link types, where C, S and D links are blue, green and red, respectively.

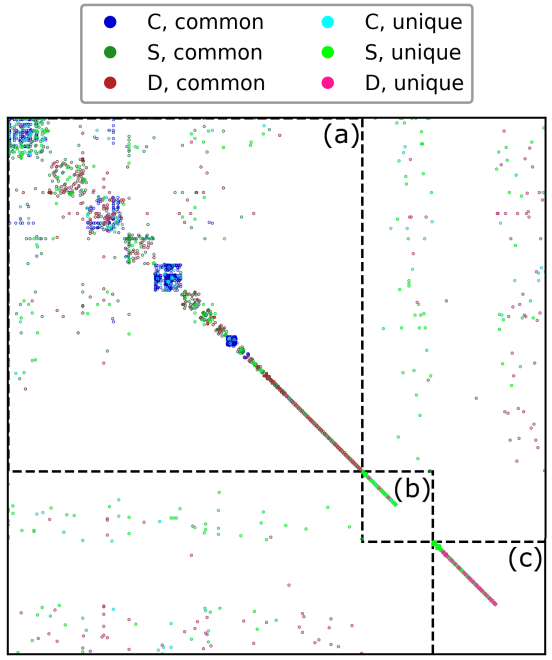


Figure 4.9: Comparison of CSD networks for bipolar disorder based on Fisher's Z transformed and weighted untransformed averages of correlation coefficients using an adjacency matrix. Nodes have first been sorted according to network origin, visualized by the dashed lines. From upper left to lower right: (a) common nodes for both CSD networks, unique nodes for (b) weighted untransformed and (c) Fisher's Z transformed averages. The nodes were subsequently ordered according to their community structure determined by the Louvain algorithm [29]. The links are coloured according to link type and network origin, where unique refers to links present in just one of the CSD networks.

Despite a good agreement between degrees in the CSD networks, it is not given that one node must have the same neighbours in both networks. This is reflected in Figure 4.8 and 4.9 where several common nodes are connected by unique links. Hence, the similarity of the neighbours for each node in the networks was investigated. Figure 4.10b illustrates the similarity between the neighbourhoods in the CSD networks using the network based on weighted untransformed correlation coefficients as reference. A similar plot using the network based on Fisher's Z transformed correlation coefficients as reference is given in Appendix E.3. Obviously, nodes that are unique to one CSD network do not share any neighbours with the second CSD network. Hence, they will have a Jaccard index of 0. As seen in Figure 4.10b, the similarity of the neighbourhoods of small-degree nodes is quite variable. Some small-degree nodes have quite low Jaccard indices, indicating that few neighbours of the given nodes are shared between the two networks. However, most small-degree nodes have a high similarity. For high-degree nodes, the similarity is also generally high and all nodes with degrees above 10 have Jaccard indices above 0.38. It is noteworthy that the neighbourhood of the top hub in the CSD network based on weighted untransformed correlation coefficients only has a Jaccard index of 0.53 when compared to

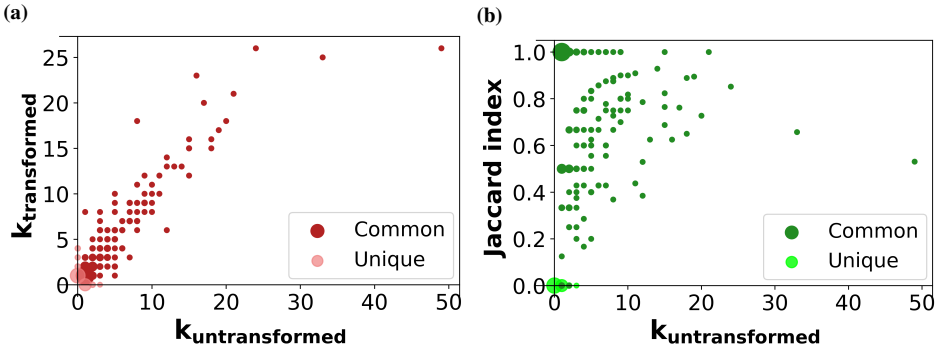


Figure 4.10: Comparison of (a) degrees and (b) neighbourhoods in the CSD networks for bipolar disorder based on Fisher's Z transformed and weighted untransformed averages of correlation coefficients. The neighbourhood analysis uses the CSD network based on weighted untransformed averages as reference. k denotes the node degree in the CSD network indicated by its subscript. The terms "common" and "unique" refer to whether the nodes are shared between the two CSD networks or is unique to one of them. The size of the points in the plots reflects the number of nodes with the given characteristic.

the CSD network based on Fisher's Z transformed correlation coefficients. This is probably due to the differences in the degree of this hub between the networks.

To sum up, the two methods for generating consensus CSD networks produce quite similar networks for BP, with comparable degrees for common nodes. The similarity is greatest for the C networks, but also good for the S and D networks. Importantly, the nodes that are unique for one of the combination methods are located in the periphery and have low degrees. In addition, common small-degree nodes show considerable variation of the similarity between their neighbours. High-degree nodes on the other hand, are always common for the CSD networks and have at least some of same neighbours in both networks.

4.2.4 Structural Network Analysis

Structural network analyses of the CSD networks for BP have been carried out to evaluate if these networks exhibit the characteristic features of CSD networks. Analyses of degree distribution, average clustering, assortativity and node homogeneity have been conducted and will be presented here. As the weighted untransformed averages emerged as the most suitable combination method in Section 4.2.2, the CSD network based on this measure will be in focus. The results from similar structural analyses of the CSD network based on Fisher's Z transformed averages are included in Appendix F, but are not evaluated or discussed in any further details.

The degree distribution of the CSD network based on weighted untransformed averages of correlation coefficients is displayed in Figure 4.11. The distribution is well approximated by a power law with degree exponent of 1.7, indicating a scale-free character in the net-

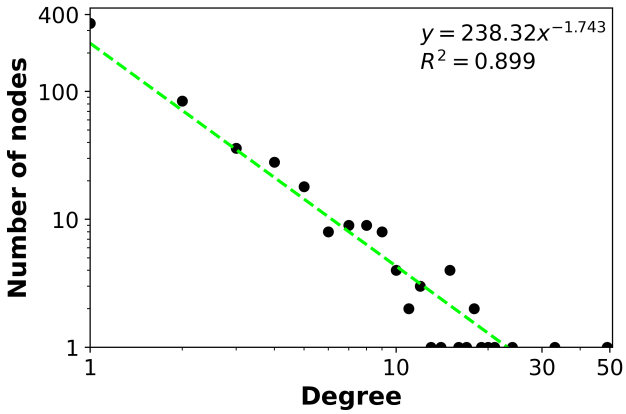


Figure 4.11: Degree distribution of the consensus CSD network for bipolar disorder based on weighted untransformed averages of correlation coefficients. The green, dashed line represents the fitted power law.

work. This illustrates that there is a co-existence of hubs and many small-degree nodes in the CSD network. The low degree exponent also indicates an important role for the hubs in the network topology.

The next structural analysis of the CSD network for BP involved calculation of degree assortativity coefficients and average clustering coefficients. This is presented in Table 4.2. The full CSD network, as well as the C network, have degree assortativity coefficients approximately equal to zero. Hence, these networks are defined as neutral and are not significantly different from random networks with identical degree sequences. However, the CSD and C networks have average clustering coefficients which are significantly higher than random expectations. This indicates that the nodes, especially those connected by C links, tend to group together in the CSD network.

The S and D networks have significant negative assortativity coefficients and are defined as disassortative with respect to degree. This means that these networks are characterized by a hub-and-spoke topology. This trend is more pronounced for the D than the S network. Furthermore, both networks have average clustering coefficients equal or close to zero. This means that few neighbours of a node in the S network, and none in the D network, link to each other. As indicated in Table 4.2, the average clustering coefficients are not significantly different from random networks with identical degree sequence. This suggests that the low clustering coefficients may result from the degree sequence itself.

The final structural analysis focused on homogeneity and evaluation of the tendency of nodes to be connected with different link types. Figure 4.12a illustrates that 99 of the 566 nodes in the CSD network based on weighted untransformed averages of correlation coefficients, are connected to other nodes through two different link types. In contrast, only four nodes have interactions of all types (C, S and D). A further exploration of the homogeneity as a function of degree in the CSD network is given by the box plot in Figure

Table 4.2: Assortativity and average clustering in the CSD network for bipolar disorder based on weighted untransformed averages of correlation coefficients. *P* values are calculated from expectations from random networks with identical degree sequence as the indicated network. Significant *P* values are marked in red and the directions of significant differences are indicated.

Network type	Degree assortativity			Average clustering		
	Coefficient	Adjusted *P* value	Compared to expectation	Coefficient	Adjusted *P* value	Compared to expectation
Full CSD network	-0.0065	> 0.05	–	0.094	< 0.05	Higher
C network	-0.039	> 0.05	–	0.41	< 0.05	Higher
S network	-0.15	< 0.05	Lower	0.015	> 0.05	–
D network	-0.22	< 0.05	Lower	0	> 0.05	–

4.12b. In most cases, high-degree nodes ($k \geq 10$) have high homogeneity scores. This indicates that they are dominated by one type of interaction. Intermediate-degree ($3 \leq k \leq 9$) nodes exhibit more variation in their homogeneity scores. This is expected as there is a higher number of nodes within these categories. The mean homogeneity scores tend to be lower for intermediate-degree nodes than high-degree nodes, but this difference is not significant (t-test: $P = 0.059$). However, there is a significant correlation between degree and homogeneity for genes with $k \geq 3$ ($r_s = 0.17, P = 0.020$). Genes with $k \leq 2$ have been omitted from this evaluation as they can have a maximum of one or two interaction types. In particular, a node with degree of one will always have a homogeneity score of one. For a node with degree of two, the homogeneity score must be 0.5 or one.

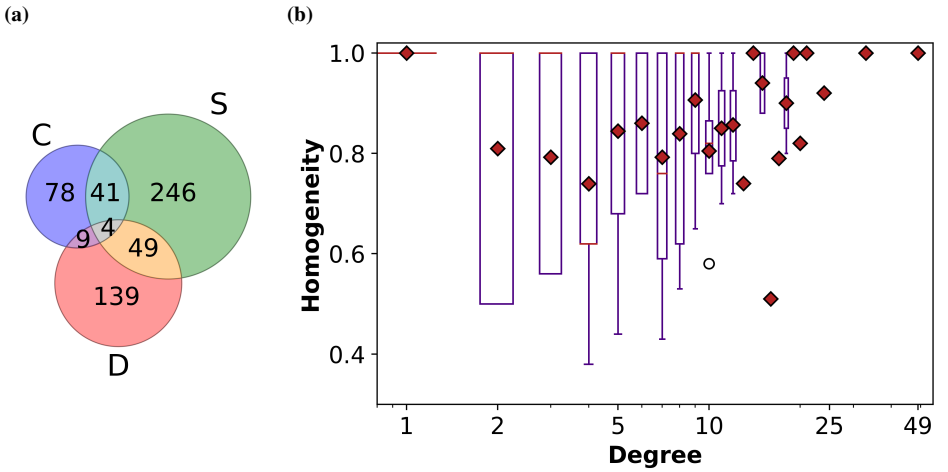


Figure 4.12: (a) Number of nodes involved in each type of interaction and (b) node homogeneity scores in the CSD network for bipolar disorder based on weighted untransformed averages of correlation coefficients. Red bars and red diamonds indicate median and mean homogeneity scores, respectively.

Taken together, the structural analyses show that the degree distribution of the CSD network for BP is well approximated by a power law. Furthermore, the full CSD and C network are neutral, but have relatively high average clustering. S and D networks on the other hand, are disassortative with low average clustering coefficients. Finally, homogeneity analysis indicated that hubs are mainly dominated by one interaction type and few nodes have interactions of all types.

4.2.5 Functional Analyses of the CSD Network

The structural analyses of the CSD networks revealed important information about the network topology. However, it does not provide clues about the biological interpretation of the network. Consequently, additional functional investigations of the networks have been performed. The aim is to reveal genes and interactions that can provide information about the biological underpinnings of BP. These analyses have been restricted to the CSD network based on weighted untransformed averages due to its emergence as the most suitable combination method in Section 4.2.2.

Enrichment analyses

The CSD network fro BP, as well as the individual C, S and D networks, were subjected to GO enrichment analysis of biological processes. Surprisingly, none of the networks were significantly enriched for biological processes when requiring FDR < 0.05. It should be noted that eighth IDs from the reference set and two IDs from the CSD network were registered as unmapped and omitted by PANTHER [61, 75, 76] in the enrichment analysis.

The analysis tool from PANTHER [61, 75, 76] investigates both under- and overrepresented GO terms in the input list relative to the reference lists. However, only overrepresented GO terms are of interest here. Hence, the requirement of FDR < 0.05 may be too strict. The analyses were therefore repeated with FDR < 0.1. This still yielded no significant results for the full CSD network nor the S and D networks. Hence, the GO analysis of the CSD network does not provide clues about potential pathways that are dysregulated in BP. The C network on the other hand, was enriched for proton transmembrane transport (FDR $= 0.09$) with a fold enrichment of 7.93. Transport of protons across a membrane is vital for many cell functions, including generation of proton gradients which are important for energy transduction. Even though this function is not restricted to the brain, it is not surprising that proton transmembrane transport emerges as an enriched GO term in the C network.

The GO enrichment analysis was followed by a disease enrichment analysis of the full CSD network. At this point, the analysis was restricted to curated information from DisGeNET [77, 78] and only overrepresentation of genes related to BP was investigated. This enrichment analysis indicated that 29 genes in the CSD network ($P = 0.007$) have previously been associated with BP. It is noteworthy that only 363 (of 566) nodes in the CSD network were mapped to the DisGeNET database (if Entrez IDs were used as input, only 361 nodes were mapped to the database). Hence, it cannot be ruled out that additional genes in the network have been associated with BP in other studies. Nonetheless,

an enrichment of BP genes in the CSD network supports a biological interpretation of the network as relevant for BP. A table with degrees and homogeneity scores for all of the 29 genes is included in Appendix G.1. It is noteworthy that most of these genes are dominated by S or D links.

Table 4.3 reports the top genes in the CSD network which have previously been associated with BP according to DisGeNET [77, 78]. In this case, the top genes are defined as genes with degrees above five. Surprisingly, four of these genes (AGT, MLC1, S1PR1, NR2E1) are mainly connected by C links. This indicates that their co-expressions with their neighbours are mainly conserved between BP and control samples. Thus, their involvement in BP is not readily apparent from their gene co-expressions in the CSD network.

The fifth and final node with a degree above five which has been associated with BP in previous studies, is DRD4 (Dopamine Receptor D4). DRD4 encodes a G-protein coupled receptor which, upon binding of dopamine, triggers intracellular signaling. Dopamine mediates a variety of functions, including reward, sleep regulation and cognitive functions [88]. These functions appear to be related to typical symptoms of BP (see Section 2.1). In the CSD network, DRD4 has mainly D links, but also one S link. This indicates a dysregulation of DRD4, thus supporting the involvement of this gene in BP. Interestingly, Zhao et al. [89] found an association between a single nucleotide polymorphism (SNP) variant in the promotor of DRD4 and BP I patients. It was suggested that this SNP could perturb transcription factor binding sites in DRD4 and is associated with DNA methylation of DRD4 [89]. This may explain the dysregulation of DRD4 observed in the CSD network. At the same time, Zhao et al. [89] claim that DRD4 may have a general role in symptoms associated with a variety of disorders and is not necessarily unique to BP.

In summary, the GO enrichment analysis did not highlight any specific biological process as dysregulated in BP. However, the CSD network is enriched for genes that have been associated with BP in previous studies. In particular, the dysregulation of DRD4 in the CSD network supports its potential role in BP.

Table 4.3: Top genes with degrees above five, as well as their degrees and homogeneity scores, from disease enrichment of the CSD network for bipolar disorder. The rows are coloured according to the main interaction type of the nodes, where C and D types are coloured blue and red, respectively. S is not the main interaction type for any of these genes.

Node	k	$\mathbf{k}_C$	$\mathbf{k}_S$	$\mathbf{k}_D$	H
AGT	18	18	0	0	1.0
MLC1	15	14	1	0	0.88
S1PR1	12	11	1	0	0.85
NR2E1	9	8	1	0	0.8
DRD4	6	0	1	5	0.72

Functional Analysis of Communities

The enrichment analyses of the CSD network were followed by identification of communities in the network. A community refers to a group of nodes in a network that are more likely to be connected to each other than to other nodes in the network. The communities in the CSD network were identified with the Louvain algorithm [29]. This revealed 15 communities which contained more than five nodes. A visual representation of these communities is provided in Figure 4.13.

The identified communities were subsequently subjected to GO enrichment analysis of biological processes. 11 of the communities showed no overrepresentation of GO terms. Three communities (community number 5, 8 and 13 in Figure 4.13) showed overrepresentation with FDR < 0.05. One additional community (number 10 in Figure 4.13) was enriched for biological processes when the requirement was adjusted to FDR < 0.1. Table 4.4 lists the enriched GO terms for each of these communities. In this table, related classes in an ontology are represented by the most specific subclass. Complete lists of all enriched GO terms in the four communities are provided in Appendix G.2.

Table 4.4: Enriched Gene Ontology (GO) terms in the communities of the CSD network for bipolar disorder. The community numbers correspond to the numeration in Figure 4.13. Only enriched GO terms with false discovery rate (FDR) < 0.05 are listed for communities 5, 8 and 13. For community 10, this requirement has been adjusted to FDR < 0.1. The presented GO terms correspond to the most specific subclass if related GO terms are enriched for a community.

Community number	**GO biological processes**	**Fold enrichment**	**FDR**
5	Proton transmembrane transport	15.51	0.0253
	ATP metabolic process	11.63	0.0411
8	Complement activation. classical pathway	>100	0.0112
	Vertebrate eye-specific patterning	>100	0.0271
	Complement-mediated synapse pruning	>100	0.0325
	Neuron remodeling	>100	0.0348
	Innate immune response	13.86	0.0369
10	Regulation of bone remodeling	23	0.0939
	Regulation of localization	7.05	0.0824
	Cell migration	4.22	0.0908
	Generation of neurons	4.09	0.0611
13	Myelination	44.86	0.0132

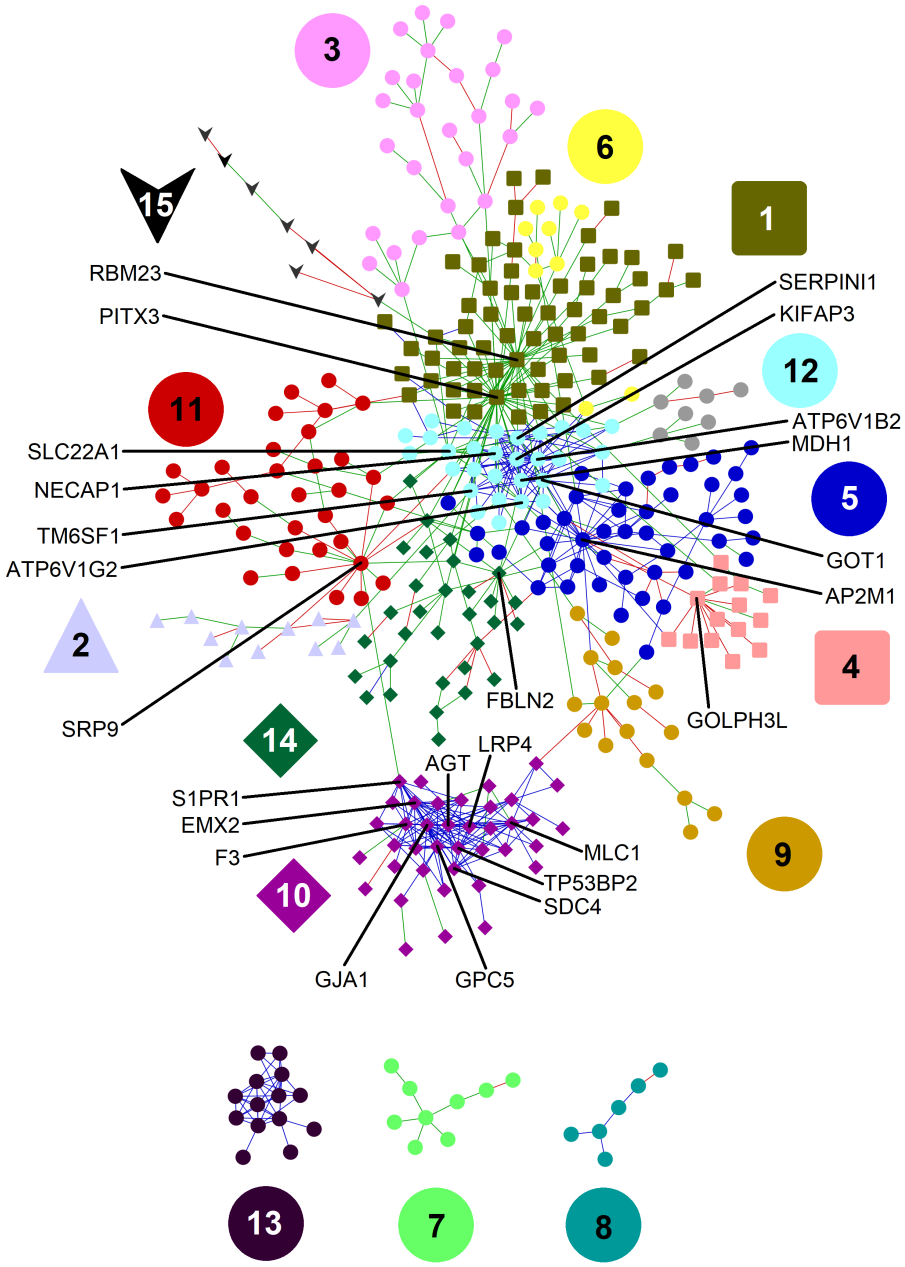


Figure 4.13: Visualization of communities and hubs in the CSD network for bipolar disorder based on weighted untransformed averages of correlation coefficients. Components with five or fewer nodes have been excluded for illustration purposes. The colour and shape of the nodes reflect their community membership. Numbers represent numeration of the communities. Links are coloured according to their interaction type, where C, S and D links are blue, green and red, respectively. In addition, nodes with degrees above nine are labelled for emphasis.

It is noteworthy that communities which showed overrepresentation of GO terms mainly consist of nodes connected by C links. This means that most of the co-expressions of these genes are conserved between BP and control samples. It is reassuring that many of these enriched GO terms are related to nervous functions as data from brain tissue has been analyzed. The remaining terms include proton transmembrane transport, ATP metabolic process, regulation of localization, cell migration and terms related to immune responses. It is not surprising to find enrichment of these terms as many represent general and important functions for all cells/tissues. It is however, more surprising that the communities are also enriched for vertebrate eye-specific patterning and regulation of bone remodeling, given that the samples originate from the DLPFC. This could simply be caused by overlap of genes involved in the GO terms. For instance, the two genes (C3, C1QA) annotated to vertebrate eye-specific patterning are also annotated to all other GO terms for community 8 in Table 4.4. A similar overlap may cause genes related to regulation of bone remodeling to become overrepresented in community 10. It is also noteworthy that the FDR requirement has been adjusted to 0.1 for community 10, making it more likely to have a false positive among the overrepresented GO terms for this community.

As mentioned above, all of the communities with enriched GO terms consist mainly of C links (or exclusively C links in the case of community 13). In general, S and D links are of more interest as they may provide clues about potential disease mechanisms. Consequently, the following paragraphs are devoted to genes with mainly S and D links in the communities which showed enrichment of GO terms, starting with community 5 (Figure 4.13). In this community, the two nodes C2orf42 (Chromosome 2 Open Reading Frame 42) and RPA2 (Replication protein A 32 kDa subunit) have only D and S links. This may indicate a dysregulation of C2orf42 and RPA2 with respect to the other community members. C2orf42 is an uncharacterized protein, and its potential role in BP is thus unknown. RPA2 functions as a subunit of the replication protein A (RPA) complex which binds single-stranded DNA [90]. This complex is involved in many aspects related to DNA metabolism, including DNA replication, recombination and repair, as well as cellular response to DNA damage as a result of stress [90]. A previous study [91] found that RPA2 is a hub in the interaction network of valproate, a medication commonly used to stabilize mood in BP. It is known that some of the patients in the studies included in the consensus CSD network for BP have been treated with valproate [71, 72]. Thus, the suggested dysregulation of RPA2 in the CSD network may reflect valproate treatment.

Interestingly, community 10 (Figure 4.13) is only connected to the giant component of the CSD network through one S and one D link. These S and D links connect S1PR1 and PHGDH (members of community 10) to SRP9 and SRP14, respectively. Furthermore, both SRP9 and SRP14 have only S and D links to other nodes in the CSD network. Based on transitivity of strong correlations, it is also expected that SRP9 and SRP14 exhibit specific or differential co-expression patterns to the neighbours of S1PR1 and PHGDH (as they are connected by C links). However, these patterns have not been detected in the CSD network under the given importance and significance levels. Nevertheless, this may indicate a dysregulation of SRP9 and SRP14 in BP. Interestingly, SRP9 and SRP14 are known to encode proteins that, together with the Alu portion of the signal recognition par-

ticle (SRP) RNA, form the elongation arrest domain of the SRP complex [92]. The SRP complex is a ribonucleoprotein complex with a role in targeting specific proteins to the endoplasmic reticulum (ER). The elongation arrest domain is required to cause a delay of ribosomal elongation of nascent peptide chains. This delay allows delivery of proteins to correct compartments during their biosynthesis [92]. Faoro and Ataide [93] suggest that "*the glutamate network might be a key target of SRP in the brain*". This fits well with a potential role for SRP9 and SRP14 in BP, given that the glutamatergic system has been suggested to be involved in BP pathophysiology [12]. However, the individual SRP components have also been related to new roles outside the SRP complex, where SRP9 and SRP14 have been shown to function as regulators of translation and in stress response [93]. The multiple functions of SRP9 and SRP14 make it difficult to draw firm conclusions about their specific contributions to BP.

Taken together, the GO enrichment of community number 5, 8, 10 and 13 suggests a conservation of many important biological functions between BP and control samples. These are related to specific processes in the nervous system, but also to more general terms that are important for many cell types and tissues. The analysis has also suggested potential roles for C2orf42, RPA2, SRP9 and SRP14 in BP. However, the observed changes in co-expression patterns of RPA2 may be due to valproate treatment.

Functional Analysis of Hubs

The next functional analysis of the CSD network involved identification of hubs. Hubs refer to highly connected nodes in a network, although there is no strict degree requirement. In a CSD network, hubs have potential functional importance as they represent genes that are co-expressed with several other genes. A visual representation of all nodes with degrees above nine is included in Figure 4.13. A more detailed overview of the degree and homogeneity scores of these nodes is provided in Table 4.5. Nodes with mainly S and/or D links are of most interest as these are specific or differentiated between BP and control. Thus, these nodes may provide clues about disease mechanisms in BP. Consequently, the following analysis is limited to these genes (PITX3, RBM23, SRP9, GOLPH3L, SLC22A1 and FBLN2).

PITX3 (Pituitary homeobox 3) is the main hub in the CSD network for BP. It has a degree of 49, and all links are defined as S links. PITX3 is a transcriptional regulator important for differentiation and function of mesodiencephalic dopaminergic (mdDA) neurons [94, 95]. These neurons have a role in regulation of emotion-related behaviour and are affected in many neurological and psychiatric disorders [95]. Consequently, a dysregulation of PITX3 seems likely to be involved in the disease mechanism of BP. However, PITX3 expression is typically restricted to the midbrain dopamine neurons in the adult brain [96]. As the consensus CSD network for BP is based on samples from the DLPFC, it is thus surprising that PITX3 emerges as the largest hub in the network. This could suggest a potential alteration in the specification process in the DLPFC of BP patients. It would also have been interesting to perform a new CSD analysis of BP patients to investigate gene expression patterns in the typical brain regions for mdDA neurons (substantia nigra pars compacta, ventral tegmental area or retrorubral field [95]). It would be interesting to examine if these

Table 4.5: Degrees and homogeneity scores of nodes in the CSD network for bipolar disorder with degrees above nine. The rows are coloured according to the main interaction type of the nodes, where C, S and D types are coloured blue, green and red, respectively.

Node	k	k_C	k_S	k_D	H
PITX3	49	0	49	0	1.0
RBM23	33	0	33	0	1.0
AP2M1	24	23	0	1	0.92
GJA1	21	21	0	0	1.0
KIFAP3	20	18	2	0	0.82
TP53BP2	19	19	0	0	1.0
AGT	18	18	0	0	1.0
GOT1	18	16	2	0	0.8
MDH1	17	15	2	0	0.79
SRP9	16	0	7	9	0.51
EMX2	15	15	0	0	1.0
MLC1	15	14	1	0	0.88
ATP6V1B2	15	15	0	0	1.0
SERPINI1	15	14	1	0	0.88
GPC5	14	14	0	0	1.0
NECAP1	13	11	2	0	0.74
GOLPH3L	12	0	2	10	0.72
S1PR1	12	11	1	0	0.85
SLC22A1	12	0	12	0	1.0
LRP4	11	11	0	0	1.0
FBLN2	11	0	9	2	0.7
TM6SF1	10	7	3	0	0.58
F3	10	9	1	0	0.82
SDC4	10	10	0	0	1.0
ATP6V1G2	10	9	1	0	0.82

brain regions exhibit a similar dysregulation of PITX3 as this is expected to have the most functional consequences.

RBM23 (Probable RNA-binding protein 23) is the second largest hub in the CSD network with a degree of 33. Similar to PITX3, RBM23 is only connected to its neighbours by S links. Although RBM23 was not identified in the disease enrichment analysis of the CSD network, it has shown a suggestive association with BP in a previous GWAS study [97]. However, it should be noted that the association was only apparent in one family cohort [97]. RBM23 functions as a transcription coactivator in steroid hormone receptor-mediated transcription and as a precursor mRNA (pre-mRNA) splicing factor [98]. In contrast to related constitutive splicing factors, RBM23 is believed to have a role in regulation [98]. Consequently, a potential dysregulation of RBM23 is expected to have consequences for regulated expression and splicing of other genes. Both of these functions are expected to influence complex traits [99]. Furthermore, alternative splicing may potentially affect protein function of its targets without altering their overall expression [99]. This suggests a role of alternative splicing in BP, a disease mechanism which has received little attention so far.

The potential disease function of SRP9 (Signal recognition particle 9 kDa protein) has

been described above, as a part of the functional investigation of community 10. Even though the potential function of SRP9 in BP will not be repeated here, it is noteworthy that SRP9 emerges as an important gene in several of the functional analyses.

GOLPH3L (Golgi phosphoprotein 3-like) is mainly connected to its neighbours in the CSD network by D links, indicating differential co-expression between BP and control samples. GOLPH3L binds to phosphatidylinositol-4-phosphate and localizes to Golgi [100]. It is believed to antagonize GOLPH3 (Golgi phosphoprotein 3), a protein involved in maintenance of the Golgi architecture as well as vesicle budding important for anterograde transport [100]. The involvement of GOLPH3L in BP seems thus plausible as its dysregulation may affect transport of receptors, ion channels and other signaling molecules. This gene has also been implicated in schizophrenia [101], a disorder which share many similarities with BP [68, 70, 73, 102]. This further supports a potential role of GOLPH3L in BP.

SLC22A1 (Solute carrier family 22 member 1) has a degree of 12 in the CSD network and is connected to all its neighbours by S links. SCL22A1 encodes a transmembrane protein which transports organic cations [103]. It is primarily expressed in the liver [103], but its messenger RNA (mRNA) has also been detected in the brain [104]. In the brain, SCL22A1 is believed to be involved in translocation of cations, such as norepinephrine, serotonin, dopamine, acetylcholine and histamine, across the blood-brain barrier [104]. It is thus tempting to suggest that a potential dysregulation of SCL22A1 in BP, as indicated by the CSD network, may disrupt flux of organic cations across the blood-brain barrier. However, the biomedical relevance is still unclear due to overlapping functions of related cation transporters [104]. Moreover, several antipsychotic and antidepressant administered to BP patients, at least in the data set GSE12469 [72], act as substrates and/or inhibitors of the protein encoded by SCL22A1 [104]. Consequently, the indicated dysregulation of SCL22A1 in the CSD network may reflect medication rather than involvement in the disease mechanisms for BP.

The final node that will be investigated here is FBLN2 (Fibulin-2), which has a degree of 11 in the CSD network. FBLN2 encodes a glycoprotein which is secreted to become a constituent of the extracellular matrix in basement membranes, elastic fibers and other connective tissue structures [105]. Consequently, FBLN2 is believed to be involved in the formation of scaffolds for cells and tissues [105]. In addition, FBLN2 has been identified as a key mediator of pro-neurogenetic effects on neural stem cell (NSC) via TGF-β1 [106]. As a result, a dysregulation of FBLN2 in BP may potentially affect neurogenesis. On the other hand, knockouts of FBLN2 in mice produce viable, fertile and anatomical normal mice [105]. Thus, the consequences of a potential dysregulation of FBLN2 in BP are unclear and are possibly compensated for by other members of the fibulin family.

To sum up, the major hubs with S and D links are generally involved in quite broad processes and their gene products affect many downstream proteins. PITX3 and RBM23 are involved in transcription, and RBM23 is also involved in alternative splicing. SRP9 and GOLPH3L are involved in transport and localization of proteins, while SLC22A1 has

a role in transport of several organic cations. FBLN2 is an extracellular matrix protein which has also been shown to affect neurogenesis.

Comparison with Basal Ganglia

In addition to investigations of the biological function of hubs, a manual investigation of their underlying correlations for S and D links was also conducted. For the D links of SRP9 and GOLPH3L, there is a shift from negative correlations in control samples to positive correlations in BP samples. For all investigated S links of the hubs, there are generally strong correlations for the gene pairs in BP samples and insignificant correlations in control samples. For PITX3, SLC22A1 and FBLN2, the S links are mainly negatively correlated in BP samples. For RBM23, SRP9 and GOLPH3L, the S links are mainly positively correlated in BP samples. In any case, this indicates a gain of correlation for several gene pairs in BP. An extension of this analysis to all S links indicated that this was a general finding where 282 S links represent a gain of correlation in BP. In contrast, only 18 S links represent a lost correlation in BP.

The emergence of strong correlations in the BP samples compared to insignificant correlations in the control samples could indicate an alteration of the specification process of the DLPFC in BP samples. This hypothesis is further supported by the emergence of PITX3 and FBLN2 as hubs in the CSD network. An alteration of the specification process could shift the DLPFC in BP samples in the direction of other brain regions relative to the control DLPFC. Consequently, the underlying correlations of the CSD network (based on weighted untransformed averages) were compared with correlations from basal ganglia. There is no reason to believe that the BP DLPFC has been shifted towards basal ganglia in particular, but this brain region was chosen due to the availability of relevant data from [9].

A total of 747 links were included in the CSD network for BP (with $p = 10^{-4}$, followed by significance filtering). However, only 727 of these links have been evaluated in the data set from basal ganglia. Hence, the comparison will be based on these 727 links. An overview of the comparisons is provided in Table 4.6. Interestingly, Table 4.6 indicates a shift in BP DLPFC towards basal ganglia. This is true for all link types, both individually and combined. This supports the hypothesis of a potential alteration in the specification process of the DLPFC in BP.

Table 4.6: Comparison of correlations from all gene pairs in the CSD network for bipolar disorder (BP) to corresponding correlations in basal ganglia. For each gene pair, the potential shift of correlations from the dorsolateral prefrontal cortex of BP samples towards correlations from basal ganglia has been evaluated.

Link type	BP shifted towards basal ganglia		BP not shifted towards basal ganglia		Adjusted *P* value
	Count	Percentage [%]	Count	Percentage [%]	
C	231	75	76	25	0.0052
S	221	75	72	25	0.0050
D	110	87	17	13	$1.6{*}10^{-6}$
CSD	562	77	165	23	$1.7{*}10^{-9}$

Chapter 5

Discussion

The major aim of this thesis is to investigate differences in gene co-expressions between BP and control samples. Previous studies of BP have typically suffered from small sample sizes [13], as illustrated by the six data sets [68–73] included in this thesis. Gordovez and McMahon [13] claimed that "*meta-analysis of multiple independent samples have perhaps the best likelihood of success*" to resolve this issue. This prompted the second aim of this thesis: to develop and test methods for constructing consensus CSD networks. In addition to specifically increase the sample sizes for BP, this would also generate an improvement of the CSD approach when dealing with small sample sizes in general. The following discussion will be divided into three main sections, where the first evaluates and compares the combination methods underlying the creation of consensus CSD networks. The second section will discuss the structural consequences of using combined correlations in CSD analyses. Finally, the last section will describe the functional findings, as well as limitations, of the consensus CSD approach for investigation of BP.

5.1 Combining Correlation Coefficients

Two methods for creating consensus CSD networks have been tested in this thesis. Both of these approaches rely on the combination of individual Spearman rank correlation coefficients into averaged values, which subsequently are used as input for the calculation of C, S and D scores in the CSD approach. The first method relies on Fisher's Z transformed averages of correlation coefficients, while the second method relies on weighted untransformed averages. From the method development, using a data set called "*Skin - Not Sun Exposed (Suprapubic)*", both methods clearly outperformed subgroup correlations as estimators of the reference correlation coefficients. However, the method development did not point out any combination method as superior to the other. Consequently, both combination methods were used to generate combined correlation coefficients for BP and control samples in the second part of this thesis. In this case, some important distinctions between the methods emerged.

The combination of correlation coefficients from BP and control samples showed that Fisher's Z transformed averages tend to generate more extreme estimates than the weighted untransformed averages. This is in line with previous studies which have illustrated that Fisher's Z transformed averages introduce a positive bias while weighted untransformed averages generate a negative bias in the estimates of reference correlation coefficients [45, p. 207, 48, 107, 108]. Hence, weighted untransformed averages can be viewed as a more

conservative method for estimating combined correlation coefficients. This is generally preferable in network analyses as it is expected to give a lower number of false positives.

The combination of correlation coefficients from BP and control samples further illustrated that the Fisher's Z transformed averages could be hijacked by spurious perfect correlations. Consequently, weighted untransformed averages appeared to be most suitable for CSD analysis. It was attempted to circumvent the influence of spurious perfect correlations in the Fisher's Z transformed averages by setting Fisher's Zs equal to 5 for these values (or -5 for perfect negative correlations). This appeared to be sufficient in the method development. However, the analysis of BP and control samples indicated that this simplification was inadequate. This is most likely due to different number of included studies/subgroups in the method development and BP analysis. If it is still wishful to use Fisher's Z transformed averages, the spurious perfect correlations could probably have been avoided by filtering out low-count genes at an earlier processing step. This step is generally recommended for analysis of RNA-Seq data [86], but has been omitted in this thesis due to the inclusion of several studies. As a gene with low expression in one data set may have a biologically relevant expression in another data set, it could be relevant for the disease/condition in question. Hence, these genes have been included in the analyses in this thesis. A similar procedure is applied in Co-expression Differential Network Analysis (CoDiNA) [109], a method for comparison of multiple networks.

In this thesis, the calculation of combined correlation coefficients using Fisher's Z transformation assumes a fixed-effect model. This means that it has been assumed that the "true" correlation coefficients in the population are constant and equal across the included data sets. This assumption might be justifiable for the method development, as the subgroups originate from the same underlying data set. However, there is no guarantee that the same assumption is applicable to the BP and control data sets as these originate from different studies. Furthermore, no tests for homogeneity of correlations have been performed. Hence, it is possible that the fixed-effect assumption is invalid for the BP and control samples. In fact, heterogeneous-effect models are generally more representative models for real-world data [110]. The weighted untransformed averages of correlation coefficients on the other hand, have been categorized as a random-effect model [48]. This means that weighted untransformed averages allow the "true" correlation coefficient to vary between data sets [48]. These theoretical considerations support the interpretation of weighted untransformed averages as superior to the fixed-effect version of Fisher's Z transformed averages for application in CSD analysis.

There are several possible reasons for the emergence of differences between the combination methods for BP and control samples and the absence of these differences in the skin data set ("*Skin - Not Sun Exposed (Suprapubic)*") utilized for the method development. This might be due to different number of investigated gene pairs, different number of total samples or different number of included studies/subgroups. Another possibility is that this discrepancy arises from differences in the underlying data sets, including differences in variation or technical differences. As the combined BP and control data set originate from several different studies, these might be affected by technical differences such as the

RNA extraction methods, expression platforms and RNA library preparation procedures [22, 23]. The 1000-gene skin data sets, used in the method development, originated from the same underlying data set (which was divided into subgroups to allow re-combination). Hence, there are no technical differences between the subgroups of this data set.

Even though the weighted untransformed averages emerged as a more suitable method than the Fisher's Z transformed averages for CSD analysis, both methods appeared to produce comparable estimates for the reference correlations in the method development. The analysis of combined correlations of BP and control samples cannot reveal which of the combination methods that is closest to the "true" values, as these "true" values are unknown. Other studies have also compared the combination methods, but with varying results. Silver and Dunlap [107] found that the method based on Fisher's Z transformed averages is less biased than weighted untransformed means. They have later been criticized for ignoring the effect of number of studies. Consequently, their conclusion is expected to only apply to analysis of a small number of studies [108]. Furthermore, Strube [108] noted that the size of the biases of the two combination methods approaches each other as the number of studies increases. This fits well with the observed similarity between the combination methods during the method development in this thesis. In contrast, Schmidt and Hunter [48] advocate the use of weighted untransformed correlations as they claim that the positive bias introduced by Fisher's Z transformation "*is always greater in absolute value than the bias in the untransformed correlation*". Hence, the choice of combination method in terms of size of the bias is still unclear. It is also unclear how many subgroups that must be combined, and how large the summed sample size must be, in order to generate acceptable estimates of the reference correlations.

In this thesis, it has been chosen to focus on combination of correlation coefficients as a method for creating consensus CSD networks. One could imagine alternative strategies that combine data sets at the level of raw data or the final C, S and D scores. However, it was chosen to combine Spearman rank correlation coefficients due to the non-parametric and generality of this measure. As outlined by Voigt et al. [9], the non-parametric nature of Spearman rank correlation coefficients makes it unnecessary for expression values emerging from different sources to be normalized against each other. Consequently, it is justifiable to combine Spearman rank correlation coefficients, without comprehensive normalization, from different studies which may have used different approaches for estimating gene expression values. This will also allow the creation of consensus CSD networks to be standardized and easily applied to other diseases or conditions.

The use of correlation coefficients for creation of networks has been criticized for inclusion of many false positive associations [23]. Hence, wTO has been suggested as a replacement for correlation coefficients in network studies [21–23]. Furthermore, it has been shown that wTO improves the fidelity of co-expression networks when dealing with small sample sizes [24]. Nonetheless, it was chosen to focus on correlation coefficients in this thesis due to the reliance of the CSD approach on Spearman rank correlation coefficients. As the sample size will increase when estimating combined correlation coefficients, it is also believed that the expected positive effects of wTO will be reduced [24].

In conclusion, combination of Spearman rank correlation coefficients as estimates of the ”true” correlation coefficients outperforms the current method for dealing with small sample sizes in CSD. Weighted untransformed averages of correlation coefficients appear to be most suitable for CSD analysis. This method is more conservative and less affected by spurious perfect correlations than Fisher's Z transformed averages. There are still some unanswered questions related to the size of bias in the combination methods, requirements for total sample sizes and requirement for number of data sets to be combined.

5.2 Structural Evaluation of Consensus CSD Network for Bipolar Disorder

The combined correlation coefficients for BP and control samples were used as input to the CSD approach to create consensus CSD networks for BP. This illustrated that the two combination methods generate quite similar networks, especially for the conserved (C) links and nodes connected by them. The two CSD networks exhibited most variability in identities and neighbourhoods for small-degree nodes. Hubs on the other hand, had similar identities, degrees and neighbourhoods in both CSD networks. This means that the same hubs are generated with both combination methods, indicating that the choice of method may have low impact on the most important nodes in the final CSD network. Nevertheless, the remainder of this discussion will be limited to the CSD network based on weighted untransformed averages as this is believed to be the most suitable method for CSD analysis. Hence, the term ”CSD network for BP” will from this point refer specifically to the network based on weighted untransformed averages.

The structural analysis of the CSD network for BP indicated that its degree distribution is well approximated by a power law. This indicates that the network topology is quite different from random Erdős-Rényi networks, which have binomial degree distributions. According to Barabási [7], most real networks approximate a power law with a degree exponent between 2 and 3. Noticeable, the degree exponent (1.7) of the fitted power law in the CSD network for BP is lower than typical degree exponents in real networks. Barabási claims that ”*the number of links connected to the largest hub grows faster than the size of the network*” [7] when the degree exponent is lower than 2. Thus, these networks cannot exist [7]. It is important to note that this argument is based on networks where the number of nodes approaches infinity. In contrast, the CSD network for BP is of finite size with 566 nodes. Consequently, the low degree exponent indicates an important role for the hubs in the network rather than an impossibility of the network topology.

The structural analyses of the CSD network for BP also indicated that the network has similar structural characteristics as conventional CSD networks. Specifically, the homogeneity assessment of the CSD network for BP indicated that homogeneity is correlated with degree, a typical characteristic for CSD networks [9]. In addition, the S and D networks for BP behave as expected from previous CSD analyses in terms of low average clustering coefficients and disassortative characteristics [9]. This means that these networks are characterized by hub-and-spoke topologies, where few neighbours of a node, or

none in the D network, are directly connected to each other. Moreover, the C network has a relative high average clustering coefficient as expected for this subnetwork type [9]. This means that nodes connected by C links tend to group together in the network. Combined, this illustrates that weighted untransformed averages of correlation coefficients seem suitable as input for the construction of CSD networks. Conversely, the C network for BP has a degree assortativity coefficient approximately equal to 0 and is defined as neutral. Typically, C networks are defined as assortative [9]. Voigt et al. [9] argued that this assortativity is a natural consequence of transitivity of strong correlations. Specifically, gene A tends to be connected to the neighbours of gene B if gene A and B are strongly correlated. Hence, gene A and B are expected to share a similar number of neighbours [9]. The absence of this finding in the CSD network for BP may simply be a consequence of the chosen importance and significance levels in this thesis. It is possible that the assortative character of the C network would emerge if these requirements were relaxed.

In contrast to typical CSD analyses, the subsampling algorithm and calculation of variances have been omitted in the construction of consensus CSD networks for BP. There are three main reasons for this choice. First, this thesis has been restricted to testing methods for generating combined correlation coefficients. Thus, the investigation of possible methods for generating combined variance scores is beyond its scope. Second, it is recommended to omit subsampling and calculation of variances when dealing with small sample sizes in CSD analyses [9]. Consequently, both the current CSD approach and the newly developed consensus CSD networks lack calculation of variances when dealing with small sample sizes. As a result, the consensus CSD network is expected to be an improved representation of the condition in question as the combined correlation coefficients emerged as better estimates of the ”true” correlation coefficients in this thesis. Third, a previous master thesis [111] has shown that omitting calculation of variances has minor impact on the robustness of identifying disease-related genes. Taken together, this justifies the exclusion of variance calculations in the consensus CSD analysis.

In this thesis, an additional step with significance filtering was introduced in the construction of the CSD network for BP. The conventional CSD approach is normally only based on an importance level with the purpose of mapping the C, S and D scores to a common scale [9]. This is a weakness with the conventional CSD approach as it may allow non-significant links to be included in the final network. This is probably most important for analyses with small sample sizes or where a relatively large fraction of the input gene pairs is included in the final CSD network. It is also important for investigation of disorders, diseases or other conditions where there may be few consistent differences between condition and control.

Taken together, the structural analyses of the CSD networks indicate that the choice of combination methods may have minor impact on the final network. The CSD network for BP based on weighted untransformed averages of correlation coefficients further illustrated that this method produces networks with many typical characteristics of CSD networks.

5.3 Functional Network Analysis of Bipolar Disorder

The above discussion indicates that weighted untransformed averages of correlation coefficient may be suitable for creation of a consensus CSD network for BP. The remainder of the discussion is restricted to functional analyses of this CSD network and will investigate differences in gene co-expressions between BP and control samples. The last part of this section will describe limitations and potential sources of errors in the functional analyses.

The functional analyses of the CSD network for BP highlighted three potentially dysregulated genes (SRP9, SRP14 and GOLPH3L) involved in localization of proteins. RBM23 could also be involved in localization as it has a role in alternative splicing [98], a process which may influence subsequent mRNA location of the targets [112]. Deregulation of components in the trafficking machinery could inactivate or misregulate target proteins, or even give the mislocated proteins harmful properties [113]. Mislocalization of proteins have been associated with a range of human diseases, including Alzheimer's disease, kidney stones and cancer [113]. Consequently, its potential contribution to disease mechanisms in BP seems biologically plausible. However, it is difficult to pinpoint its exact effects as it may influence a broad range of targets.

The functional analyses of the CSD network for BP also suggested a potential alteration in the specification process of the DLPFC. This alteration is indicated at the level of individual genes, where the major hub PITX3 is known to be involved in differentiation of mdDA neurons [94, 95] and the extracellular protein FBLN2 has a suggestive role in neurogenesis [106]. It is also possible that a dysregulation of RBM23, the second major hub in the network, may affect the specification process of the DLPFC. RBM23 is believed to be involved in steroid receptor-dependent regulation [98], but its precise role is unknown. A potential alteration of the specification process of the DLPFC is also apparent at the network level. Specifically, the majority of S links in the network indicates a gain of correlations in BP relative to control. In addition, the underlying correlations in the CSD network for BP indicate a shift of BP DLPFC from control DLPFC towards basal ganglia. Changes in the specification process seem as a plausible contributor to disease mechanisms in BP as subtle alterations in developmental processes may give rise to neurological changes that become functionally significant later in life [114, 115]. Nonetheless, the explicit consequences of such alterations in BP remain to be elucidated.

A potential contribution of defects in NSC proliferation and differentiation has previously been suggested for several brain disorders, including BP [115–117]. Importantly, Chen et al. [115] studied the differences between neurons originating from induced pluripotent stem cells (iPSC) and suggested an alteration of neuronal identity in BP compared to control. The control neurons had a significantly higher expression of genes conferring dorsal telencephalic fate. In contrast, BP neurons had an increased expression of ventral determinants. These determinants are typically involved in generation of the medial ganglionic eminence and differentiation of GABAergic interneurons [115]. This supports the observations in this thesis, which suggest an alteration of the specification process in BP DLPFC.

It is possible that observations supporting the suggested alterations of specification processes in this thesis reflect other disease mechanisms in BP. For instance, the CSD network has relied on macrodissected brain tissue. Hagenauer et al. [69] claim that it is unlikely that psychiatric illnesses affect all cell types equally. As a result, it is difficult to distinguish between differences that arise due to alterations in gene expression patterns representing changes in relative number of cell types (population changes) and alterations that reflect dysregulation of cells at the individual level [69]. Consequently, it is possible that the functional analyses reflect changes in cell type proportions rather than changes in the specification process. Specifically, the shift of BP DLPFC towards basal ganglia could indicate that the cell type proportions in BP DLPFC are shifted towards proportions in basal ganglia. Changes in cell type proportions could be related to the disease mechanisms for BP, as suggested by Ramaker et al. [68], or be caused by confounding factors. As an example, prolonged hypoxia may give low pH which could alter cell type proportions. At the same time, it is important to note that pH can be caused by other factors than hypoxia, including BP itself [69]. As a result, it is challenging to disentangle which changes in cell type proportions that are caused by BP and which are caused by other factors.

The comparison of BP DLPFC to basal ganglia requires some additional attention. Initially, it may be surprising that the BP correlations underlying C links in the CSD network also appear to be shifted towards basal ganglia. This observation may be explained by the following hypothetical example: imagine that the co-expression of gene A and B in BP DLPFC is shifted towards basal ganglia, but the co-expression is also conserved between control DLPFC and basal ganglia. In this case, transitivity of strong correlations implies that the co-expression must also be conserved between BP and control DLPFC. This explains the emergence of observed C links in the CSD network, even though the BP correlations may be shifted towards basal ganglia. It is also important to note that the comparison of the underlying correlations in the CSD network to basal ganglia may be affected by non-disease factors. For instance, there are smaller sample sizes underlying BP DLPFC and basal ganglia data compared to control DLPFC. As indicated by significance tests of correlations [34], high correlations are more likely to arise when analysing data sets with small samples sizes than large sizes. In some cases, this may shift BP correlations towards basal ganglia even when there is no biological reason for this shift. In addition, the design of the comparison will count cases where basal ganglia and control DLPFC are nearly identical, but small differences make BP DLPFC slightly shifted towards basal ganglia. This may exaggerate the observed number of times BP is shifted towards basal ganglia.

The disease enrichment of the CSD network for BP showed an overrepresentation of BP genes. This is reassuring as it supports a biological interpretation of the network. However, it should be noted that the disease enrichment of BP does not offer an option to provide a reference gene set. Hence, this analysis may suffer from sample source bias (discussed below). Furthermore, it is surprising that four of the top five genes identified by the disease enrichment for BP are mainly connected to other genes in the CSD network by C links. This indicates that these nodes are mainly conserved between BP and control samples. It is of course possible that these genes would have shown specific or

differentiated interactions with genes that have not been analysed in this thesis, illustrating a consequences of a relatively small number of included genes in the CSD analysis. It is also important to point out that genes may contribute to disease mechanisms at other levels than gene co-expressions. For instance, mutations in the coding region of a gene [10] or alternative splicing [99] may influence the function of a gene without altering its expression. Similarly, post-translational modifications may also contribute to disease and act on protein level [10]. As a result, many known disease genes are not registered as differentially expressed [10] and this could explain why some disease-related genes appear to be conserved in the CSD network for BP. This illustrates that co-expressions cannot capture the complete picture of a complex disease.

The functional analysis of the communities in the CSD network for BP indicated that four communities, with mainly C links, were enriched for biological processes (with FDR < 0.1). In most cases, the identified overrepresentation makes sense as the GO terms correspond to general, important functions for all cells/tissues or are directly related to nervous functions. Despite this, the emergence of myelination as overrepresented in community 13 was surprising. This community contains exclusively C links and is expected to represent a completely conserved community. In contrast, a previous study [118] has found myelination changes in BP. As the CSD network is only based on 3148 genes, it is possible that non-included myelin-related genes would have shown a dysregulation in BP. The discrepancy may also be caused by differences in medication or other confounding factors (discussed below).

The most likely explanation for the inconsistency between the observed and expected role of myelination in BP is differences in investigation methods. Importantly, Tkachev et al. [118] found that some myelin-related genes identified as significantly changed with quantitative polymerase chain reaction (PCR) were unchanged in their own microarray analyses. This was attributed to differences in dynamic range between the platforms [118], which is also applicable to the CSD analysis. In addition, Tkachev et al. [118] investigated samples from 15 BP and 15 control samples. In contrast, the analyses in thesis have been based on 139 BP and 201 control samples. Hence, the expected heterogeneity of BP [13] may have obstructed the identification of myelin-related genes as differentiated in the CSD network. Furthermore, Tkachev et al. [118] have investigated differential gene expression. The CSD analysis on the other hand, identifies changes in gene co-expressions. It is thus possible that the entire group of myelin-related genes may be up- or down-regulated in BP, while their co-expressions are still conserved. This is in line with the discussion above where it was argued that gene co-expressions cannot capture the complete picture of a disease.

The network analysis of BP has been based on samples originating from the DLPFC. This brain region was selected due to its potential role in BP. For instance, dendritic spine loss [119] and reduced activity [18, 120] have been observed in the DLPFC of BP patients. Furthermore, some studies have indicated an involvement of cellular and molecular alterations in the BP DLPFC. This includes cell growth and nervous system development alterations [121], oligodendrocyte and myelination changes [118] and mitochondrial dysfunction [72, 122]. Consequently, it was surprising that the CSD network for BP, as well

as the individual S and D networks and communities mainly consisting of S and D links, did not exhibit enrichment of any biological processes. However, this is in line with other studies that have found few or no significant changes in the DLPFC of BP patients [68, 71]. Ryan et al. [71] suggested that this could reflect that the DLPFC is weakly affected in BP.

It is also possible that the enrichment and functional analyses of the CSD network for BP have suffered from an inadequate number of included genes. A reference set of 3148 genes was evaluated in the construction of the CSD network. In contrast, about 13 300 genes are expressed in the human brain [123]. Consequently, it is possible that the CSD network misses some pathways and genes that contribute to disease mechanisms in BP as many genes have not been investigated. As an example, none of the three genes most commonly associated with BP (ANK3, CACNA1C and TRANK1) [13] were included in the reference gene set. Consequently, none of them are present in the CSD network even though they might be dysregulated in BP.

A low number of included genes in the reference gene set may also have affected the statistical power for identifying enriched biological processes with PANTHER [61, 75, 76]. Moreover, it may be statistically challenging to identify overrepresented processes, especially if there are few relevant genes in the data set for the disorder in question, as all enrichment tests should be adjusted for multiple testing. The number of reference genes for the CSD network could have been increased by relaxing the processing steps of the individual data sets, such as allowing combination of averaged expressions from one data set with possible unaveraged expressions from another data set. However, this would have come at a cost of potentially erroneous combinations. Hence, this was not conducted in this thesis. According to PANTHER [61, 75, 76], "*uploading a reference list is optional*" [60] when evaluating enrichment of biological processes. As a result, the loss of statistical power caused by a small reference set may have been circumvented by using a default reference list. However, this would have created a sample source bias in the results [124]. This means that if an input list of expressed genes in the brain is compared to a reference set with all genes expressed in the human body, it is expected that the input list will show overrepresentation of genes related to the brain. In this case, the overrepresented GO terms would merely reflect the sample source rather than the disease state [124]. Consequently, it was chosen to upload a reference list at PANTHER that corresponded to all genes evaluated in the CSD analysis to avoid such sample source bias.

The lack of overrepresented GO terms for the CSD network for BP, as well as the individual S and D networks and communities mainly consisting of S and D links, may have been caused by an inherent limitation to GO enrichment analyses. As the biological functions of all genes are not known, GO analyses suffer from an incomplete annotation of genes [125]. This is exemplified by C2orf42, an uncharacterized protein that emerged as potentially dysregulated in community 5 in the CSD network. This is an extreme case, where little is known about the functional properties of this genes. Other genes may have some known functions, but could participate in additional biological processes as well. As an example, SRP9 and SRP14 have canonical roles in the SRP complex, but have also potential roles in regulation of translation and stress response [93]. As a result, the functional

roles of some genes may have been overlooked in the GO enrichment analyses.

There are several additional explanations, and possible sources or errors, for the lack of enrichment of biological processes which also apply to all functional analyses of the CSD network. Importantly, BP is characterized by phases of mania/hypomania, depression and euthymia [12, 13]. Hence, gene co-expressions may change during the course of BP. It is difficult to capture such dynamic changes with a static model, making it more challenging to capture potentially affected biological processes in BP with the CSD approach.

The lack of enrichment of biological processes may also reflect the heterogeneity of BP as a disorder. As described in Section 2.1, BP may be divided into several subgroups [14, p. 123] and potentially several biological distinguishable disease entities [13]. To my knowledge, none of the six studies [68–73] included in the CSD analysis of BP list the specific diagnosis of their patients. Hence, it is possible that the CSD network is based on patients with different sub-types of BP. In addition, BP is known to be comorbid with other psychiatric and nonpsychiatric disorders [12, 14, p. 132-139]. This may have further occluded the CSD analysis. Despite these shortcomings, it was expected that the CSD analysis should highlight the commonalities between the patients.

Currently, several medications may be administered to BP patients. Two studies included in the CSD analysis [71, 72] list all medication administered to their patients, where most patients have taken one or more types of medications. Ideally, the analysis of BP would rely on unmedicated patients. However, this would be an unethical study design as BP may have detrimental consequences for affected individuals [12, 15]. As the effect of medication is not investigated in this thesis, it is difficult to determine if the observed effects are caused by medications rather than BP. As an example, it is possible that the potential dysregulation of RPA2 and SLC22A1 in the CSD network reflects medication. Furthermore, it is likely that medication may have affected gene co-expressions that would otherwise emerge as specific or differentiated in the CSD network. As an example, a previous study found that medication-free BP patients exhibited an up-regulation of some mitochondrial genes [72]. In contrast, medicated BP patients typically showed a global down-regulation of mitochondrial genes [72]. Consequently, pathways and biological processes that are altered in unmedicated patients, and may contribute to the disease mechanisms of BP, may have been missed in this thesis due to medication.

It is also important to remember that the CSD network for BP relies on postmortem brains. Hence, the analyses are based on snapshots of the brains and are expected to focus on changes apparent in the end stage of the disorder [117]. Information about underlying mechanisms, which may be important at earlier stages, may be missed [117]. Furthermore, postmortem brains are associated with several confounding factors, including sample pH, terminal condition (often called agonal state), post-mortem interval (PMI), age, gender and relative number of cell types [69, 72]. As an example, a prolonged agonal state is typically associated with hypoxia or acidosis in the brain. This may reduce the sample pH, which again is associated with an alteration of gene expressions patterns [72]. Prolonged hypoxia may also affect relative cell numbers in the brain, thus also affecting observed global gene

expression patterns [69]. Similarly, aging has been associated with a reduction in neuronal gene expression patterns [69]. None of these factors have been explicitly accounted for in the CSD analysis and may have influenced the final outcome. However, the importance of these factors remains unknown and matched controls have been utilized to reduce their impact.

Lastly, the construction of a consensus CSD network for BP has relied on the combination of data sets from six different studies. Unfortunately, some of these studies have received brain tissue from the same brain bank. An investigation of the available patient information suggested a potential overlap between the patients in these studies. Consequently, it is possible that some patients are included in several data sets. One could argue that the specific cells in the DLPFC samples are the biological unit of interest, and these are unique to each of the included data set. However, this ignores the hierarchical nature of biological data where two cells from one patient are expected to be more similar than two cells from two different patients [126]. As a consequence, some patients may have been given an increased importance and a higher weight when combining correlation coefficients. Furthermore, this would have lead to an exaggeration of the total sample size in the consensus CSD network. Hence, a potential dependence of the data sets increases the chance of false positive findings [126]. Anyhow, the consensus CSD network must be based on at least 76 and 136 independent BP samples and control samples, respectively, as three different brain banks have provided the brain tissue. Furthermore, the clustering analysis indicated that data sets from the same brain bank do not appear to be more similar to each other than the other data sets. As a result, all analyses have been based on all six data sets where they are treated as independent.

Taken together, the functional analyses have suggested a potential role for mislocalization of proteins and alterations in the specification process in the DLPFC of BP patients. However, this was not supported by GO enrichment analyses which indicated no enrichment of biological processes in the CSD network for BP.

Chapter 6

Conclusion and Outlook

This thesis aimed at investigating differences in gene co-expressions between BP and control samples. As most studies of BP suffer from inadequate sample sizes, it was of great interest to extend the analysis to several data sets. Consequently, this thesis has been divided into two main parts. The first part focused on developing and testing methods for combining correlation coefficients which could be used as input to the conventional CSD approach. The second part utilized the newly developed procedure to investigate differences between BP and control samples.

The method development of this thesis relied on splitting a large data set into smaller subgroups. The Spearman rank correlations for all gene pairs in all subgroup were calculated and then recombined using either Fisher's Z transformed or weighted untransformed averages. In real cases, a large data set will not be split in this way and the subgroup will correspond to independent studies. However, this step was included to establish reference Spearman rank correlation coefficients to which the estimates could be compared. Subsequently, the consensus CSD approach was used to create a network for BP by combining correlation coefficients from six different BP data sets. Together, these analyses indicated that combination of correlation coefficients outperforms the current method for dealing with small sample sizes in the CSD approach. The two combination methods produced comparable results. However, the method based on weighted untransformed averages appeared to be more suitable than Fisher's Z transformed averages for CSD analysis as it is more conservative and less affected by spurious perfect correlations.

In addition to the use of combined correlation coefficients, the consensus CSD approach differs from the conventional CSD approach in two aspects. First, the subsampling algorithm and calculation of variances are omitted in the construction of consensus CSD networks. Moreover, an additional step for significance filtering has been introduced. In general, it is recommended to include this step when utilizing the CSD approach to investigate conditions where there may be few consistent differences between condition and control or when it is wishful to include a relatively large fraction of the input gene pairs in the final network. This step may also be important for the conventional CSD analysis if the investigation is based on a data set with small sample size.

Structural analyses of the consensus CSD network for BP, based on weighted untransformed averages of correlation coefficients, indicated that this network exhibited many of the typical characteristics of CSD networks. Importantly, the degree distribution was well approximated by a power law, homogeneity was correlated with degree and the S and D

networks were disassortative and had low average clustering coefficients, while the C network had a relatively high average clustering coefficient. However, the C network did not exhibit the expected assortative character.

The consensus CSD network for BP, based on weighted untransformed averages, was subsequently subjected to functional analyses. This included identification of hubs and communities, as well as disease enrichment of the CSD network and GO enrichment analyses of the CSD network, its subnetworks (C, S and D) and communities. The functional analyses were extended to a comparison of correlations underlying the CSD network and correlations from basal ganglia. The functional analyses of BP suggested that mislocalization of proteins might contribute to disease mechanisms in BP. This was mainly based on the emergence of at least three genes involved in localization of proteins as central and potentially dysregulated in the CSD network (SRP9, SRP14, GOLPH3L and possibly RBM23). Furthermore, the functional analyses of the CSD network for BP suggested a potential alteration in the specification process of the DLPFC of BP patients. This was supported by identification of at least two hubs (PITX3, FBLN2 and possibly RBM23) with potential roles in differentiation or neurogenesis. The possible alteration in the specification process of DLPFC in BP patients was also supported at the network level as the majority of S links indicated a gain of correlation in BP and that many correlations in the BP DLPFC were shifted towards correlations from basal ganglia. Potential defects in NSC proliferation and differentiation have also been suggested by previous studies of BP. On the other hand, the roles of mislocalization and alterations of specification processes were not supported by GO analyses. This illustrates that further investigation of BP is required.

This thesis leaves some unanswered questions which may be explored in more detail in future studies. Even though weighted untransformed averages of correlation coefficients appeared to be most suitable for consensus CSD analysis, some aspects remain to be elucidated. Specifically, there are some uncertainties regarding the size of the bias for both combination methods, requirement for total sample size and requirement for number of data sets to be combined. This may be investigated by extending the method development to evaluate effects of combining correlations from different numbers of subgroups, and thus different total sample sizes. As the number of subgroups and total sample size are reduced, it is also expected that the differences between the combination methods will be revealed. This could allow an estimation of the size of bias for both combination methods by comparing the combined correlations with reference correlations.

As BP is expected to be a quite heterogeneous disorder, it would have been interesting to test the consensus CSD approach with a disease where more consistent differences between condition and control are expected. Such a disease is probably less affected by the chosen importance and significance levels. Hence, this may improve the structural analyses and verify if consensus CSD networks do exhibit the characteristic features of conventional CSD networks.

The heterogeneity of BP may also explain the lack of GO enrichment in the consensus CSD network for BP. Consequently, later studies may benefit from investigating patients with

more specific sub-diagnoses of BP as these patients may share more similarities in their gene co-expressions. However, such analyses are expected to suffer from even smaller sample sizes. Moreover, the lack of GO enrichment in the CSD network for BP may also indicate that the DLPFC is weakly affected in BP. As a result, follow-up studies may want to focus on other brain regions. As an example, it could be interesting to investigate co-expressions in the substantia nigra, ventral tegmental area or retrorubral field. These regions are important sites of mdDA neurons [95], which are expected to be affected by the main hub (PITX3) from the CSD network.

The use of the consensus CSD approach to investigate gene co-expression in BP may have suffered from the inherent dynamic changes of BP and the use of postmortem brain. It is difficult to imagine a viable method which can follow gene co-expressions in the brains of living BP patients. As an alternative, one could follow the changes in neurons generated from iPSC from BP patients as performed by Chen et al. [115] and Madison et al. [116]. This allows an evaluation of changes in gene co-expressions as the neurons differentiate as well as potential fluctuations at the terminal stage. However, it will be difficult to know which phase (mania/hypomania, depression, euthymia) the neurons exhibit at the given time point. Hence, the expected dynamics of BP will complicate the analysis of disease mechanisms for this disorder.

Finally, this thesis has been restricted to analysis of changes in gene co-expressions in BP. To create a more complete picture of complex diseases, it has been suggested to perform multi-omics analyses [127, 128]. These types of analyses integrate data from several omics, such as genomics, transcriptomics, proteomics and metabolomics [127, 128]. If only one type of omics data is used, as in the CSD analyses, one may miss relevant information and it is difficult to conclude if observed differences are causes or results of disease [128]. Hence, future studies could benefit from investigating multi-omics data for BP.

Bibliography

[1] Bianconi E, Piovesan A, Facchin F, Beraudi A, Casadei R, Frabetti F, et al. An estimation of the number of cells in the human body. Annals of Human Biology. 2013;40(6):463–471.

[2] Alberts B, Johnson A, Lewis J, Morgan D, Raff M, Keith Roberts PW, et al. Molecular biology of the cell. 6th ed. Garland Science, Taylor and Francis Group; 2018.

[3] Westerhoff HV, Palsson BO. The evolution of molecular biology into systems biology. Nature Biotechnology. 2004;22(10):1249–1252.

[4] Trewavas A. A brief history of systems biology. The Plant Cell. 2006;18(10):2420–2430.

[5] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature (London). 2001;409(6822):860–921.

[6] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001;291(5507):1304–1351.

[7] Barabási AL, et al. Network science. Cambridge University Press; 2016. Available from: http://networksciencebook.com/.

[8] Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology. 2005;4(1).

[9] Voigt A, Nowick K, Almaas E. A composite network of conserved and tissue specific gene interactions reveals possible genetic interactions in glioma. PLOS Computational Biology. 2017 09;13(9):1–34. Available from: https://doi.org/10.1371/journal.pcbi.1005739.

[10] de la Fuente A. From 'differential expression' to 'differential networking' – identification of dysfunctional regulatory networks in diseases. Trends in Genetics. 2010;26(7):326 – 333. Available from: http://www.sciencedirect.com/science/article/pii/S0168952510000879.

[11] Clark DP, Pazdernik NJ, McGehee MR. Chapter 26 - mutations and repair. In: Clark DP, Pazdernik NJ, McGehee MR, editors. Molecular biology. 3rd ed. Academic Cell; 2019. p. 832–879. Available from: https://www.sciencedirect.com/science/article/pii/B9780128132883000264.

[12] Vieta E, Berk M, Schulze TG, Carvalho AF, Suppes T, Calabrese JR, et al. Bipolar disorders. Nature Reviews Disease Primers. 2018;4(1):1–16.

[13] Gordovez FJA, McMahon FJ. The genetics of bipolar disorder. Molecular Psychiatry. 2020;25(3):544–559.

[14] Diagnostic and statistical manual of mental disorders: DSM-5. 5th ed. Washington, D.C: American Psychiatric Association; 2013.

[15] Gonda X, Pompili M, Serafini G, Montebovi F, Campi S, Dome P, et al. Suicidal behavior in bipolar disorder: epidemiology, characteristics and major risk factors. Journal of Affective Disorders. 2012;143(1-3):16–26.

[16] Merikangas KR, Jin R, He JP, Kessler RC, Lee S, Sampson NA, et al. Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. Archives of General Psychiatry. 2011;68(3):241–251.

[17] Goodwin FK, Jamison KR. Manic-depressive illness: bipolar disorders and recurrent depression. vol. 2. Oxford University Press; 2007.

[18] Langan C, McDonald C. Neurobiological trait abnormalities in bipolar disorder. Molecular Psychiatry. 2009;14(9):833–846.

[19] Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nature Reviews Genetics. 2004;5(2):101–113.

[20] Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL. Hierarchical organization of modularity in metabolic networks. Science. 2002;297(5586):1551–1555.

[21] Nowick K, Gernat T, Almaas E, Stubbs L. Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. Proceedings of the National Academy of Sciences. 2009;106(52):22358–22363.

[22] Gysi DM, Voigt A, de Miranda Fragoso T, Almaas E, Nowick K. wTO: an R package for computing weighted topological overlap and a consensus network with integrated visualization tool. BMC Bioinformatics. 2018;19(1):1–16.

[23] Berto S, Perdomo-Sabogal A, Gerighausen D, Qin J, Nowick K. A consensus network of gene regulatory factors in the human frontal lobe. Frontiers in Genetics. 2016;7:31. Available from: https://www.frontiersin.org/article/10.3389/fgene.2016.00031.

[24] Voigt A, Almaas E. Assessment of weighted topological overlap (wTO) to improve fidelity of gene co-expression networks. BMC Bioinformatics. 2019;20(1):1–11.

[25] Erdős P, Rényi A. On random graphs I. Publicationes Mathematicae, Debrecen. 1959;6:290–297.

[26] Erdős P, Rényi A. On the evolution of random graphs. Publications of the Mathematical Institute of the Hungarian Academy of Sciences. 1960;5(1):17–60.

[27] Newman ME. The structure and function of complex networks. SIAM review. 2003;45(2):167–256.

[28] Newman ME. Mixing patterns in networks. Physical Review E. 2003;67(2):026126.

[29] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 2008;2008(10):P10008.

[30] Goshtasby AA. Similarity and dissimilarity measures. In: Image registration: principles, tools and methods. London: Springer; 2012. p. 7–66. Available from: https://doi.org/10.1007/978-1-4471-2458-0_2.

[31] Song L, Langfelder P, Horvath S. Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinformatics. 2012;13(1):328.

[32] Løvås GG. Statistikk for universiteter og høgskoler. 4th ed. Oslo: Universitetsforlaget; 2018.

[33] Mukaka MM. A guide to appropriate use of correlation coefficient in medical research. Malawi Medical Journal. 2012;24(3):69–71.

[34] Zar JH. Spearman rank correlation: overview. Wiley StatsRef: Statistics Reference Online. 2014.

[35] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics. 2009;10(1):57–63.

[36] Shakya K, Ruskin HJ, Kerr G, Crane M, Becker J. Comparison of microarray preprocessing methods. In: Arabnia HR, editor. Advances in computational biology. New York, NY: Springer; 2010. p. 139–147.

[37] Hubbell E, Liu WM, Mei R. Robust estimators for expression analysis. Bioinformatics. 2002;18(12):1585–1592.

[38] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249–264.

[39] Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology. 2010;11(3):1–9.

[40] Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Briefings in Bioinformatics. 2018;19(5):776–792.

[41] Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. Biology Direct. 2009;4(1):1–10.

[42] Risso D, Schwartz K, Sherlock G, Dudoit S. GC-content normalization for RNA-Seq data. BMC Bioinformatics. 2011;12(1):1–17.

[43] Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science. 2003;302(5643):249–255. Available from: https://science.sciencemag.org/content/302/5643/249.

[44] Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, et al. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature. 2011;474(7351):380–384.

[45] Fisher RA. Statistical methods for research workers. 14th ed. Edinburgh: Oliver and Boyd; 1970.

[46] Hedges LV, Olkin I. Statistical methods for meta-analysis. Orlando: Academic Press; 1985.

[47] Rosenthal R. Meta-analytic procedures for social research. vol. 6 of Applied social research methods series. Rev. ed. Newbury Park, Calif: Sage; 1991.

[48] Schmidt FL, Hunter JE. Technical questions in meta-analysis of correlations. In: Methods of meta-analysis: correcting error and bias in research findings. 3rd ed. 55 City Road: SAGE Publications, Ltd; 2015. p. 212–242.

[49] Schmidt FL, Hunter JE. Meta-analysis of correlations corrected individually for artifacts. In: Methods of meta-analysis: correcting error and bias in research findings. 3rd ed. 55 City Road: SAGE Publications, Ltd; 2015. p. 87–164.

[50] Bard JB, Rhee SY. Ontologies in biology: design, applications and future challenges. Nature Reviews Genetics. 2004;5(3):213–222.

[51] Jaccard P. The distribution of the flora in the alpine zone. New Phytologist. 1912;11(2):37–50.

[52] Chung NC, Miasojedow B, Startek M, Gambin A. Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data. BMC Bioinformatics. 2019;20(15):1–11.

[53] Dornelas M, Gotelli NJ, McGill B, Shimadzu H, Moyes F, Sievers C, et al. Assemblage time series reveal biodiversity change but not systematic loss. Science. 2014;344(6181):296–299.

[54] Chen C, Zuo Y, Ye W, Li X, Deng Z, Ong SP. A critical review of machine learning of energy materials. Advanced Energy Materials. 2020;10(8):1903242.

[55] Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geoscientific Model Development. 2014;7(3):1247–1250.

[56] van der Velde IR, van der Werf GR, Houweling S, Maasakkers JD, Borsdorff T, Landgraf J, et al. Vast CO2 release from Australian fires in 2019–2020 constrained by satellite. Nature. 2021;597(7876):366–369.

[57] Inglis B, Schwarzenberg P, Klein K, von Rechenberg B, Darwiche S, Dailey HL. Biomechanical duality of fracture healing captured using virtual mechanical testing and validated in ovine bones. Scientific Reports. 2022;12(1):1–13.

[58] Alali Y, Harrou F, Sun Y. A proficient approach to forecast COVID-19 spread via optimized dynamic machine learning models. Scientific Reports. 2022;12(1):1–20.

[59] Salkind NJ. Encyclopedia of measurement and statistics. vol. 3. Thousand Oaks: SAGE Publications; 2006.

[60] Mi H, Muruganujan A, Huang X, Ebert D, Mills C, Guo X, et al. Protocol update for large-scale genome and gene function analysis with the PANTHER classification system (v. 14.0). Nature Protocols. 2019;14(3):703–721.

[61] Mi H, Ebert D, Muruganujan A, Mills C, Albou LP, Mushayamaha T, et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Research. 2021;49(D1):D394–D403.

[62] Rivals I, Personnaz L, Taing L, Potier MC. Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics. 2006 12;23(4):401–407. Available from: https://doi.org/10.1093/bioinformatics/btl633.

[63] Ge Y, Dudoit S, Speed TP. Resampling-based multiple testing for microarray data analysis. Test. 2003;12(1):1–77.

[64] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):289–300.

[65] Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nature Genetics. 2013;45(6):580–585.

[66] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research. 2002;30(1):207–210.

[67] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Research. 2012;41(D1):D991–D995.

[68] Ramaker RC, Bowling KM, Lasseigne BN, Hagenauer MH, Hardigan AA, Davis NS, et al. Post-mortem molecular profiling of three psychiatric disorders. Genome Medicine. 2017;9(1):1–12.

[69] Hagenauer MH, Schulmann A, Li JZ, Vawter MP, Walsh DM, Thompson RC, et al. Inference of cell type content from human brain transcriptomic datasets illuminates the effects of age, manner of death, dissection, and psychiatric diagnosis. PLOS ONE. 2018;13(7):e0200003.

[70] Lanz TA, Reinhart V, Sheehan MJ, Rizzo SJS, Bove SE, James LC, et al. Post-mortem transcriptional profiling reveals widespread increase in inflammation in schizophrenia: a comparison of prefrontal cortex, striatum, and hippocampus

among matched tetrads of controls with subjects diagnosed with schizophrenia, bipolar or major depressive disorder. Translational Psychiatry. 2019;9(1):1–13.

[71] Ryan M, Lockstone H, Huffaker S, Wayland M, Webster M, Bahn S. Gene expression analysis of bipolar disorder reveals downregulation of the ubiquitin cycle and alterations in synaptic genes. Molecular Psychiatry. 2006;11(10):965–978.

[72] Iwamoto K, Bundo M, Kato T. Altered expression of mitochondria-related genes in postmortem brains of patients with bipolar disorder or schizophrenia, as revealed by large-scale DNA microarray analysis. Human Molecular Genetics. 2005;14(2):241–253.

[73] Abdolmaleky HM, Gower AC, Wong CK, Cox JW, Zhang X, Thiagalingam A, et al. Aberrant transcriptomes and DNA methylomes define pathways that drive pathogenesis and loss of brain laterality/asymmetry in schizophrenia and bipolar disorder. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics. 2019;180(2):138–149.

[74] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research. 2003;13(11):2498–2504.

[75] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nature Genetics. 2000;25(1):25–29.

[76] Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Research. 2021;49(D1):D325–D334.

[77] Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Research. 2016;45(D1):D833–D839.

[78] Piñero J, Saüch J, Sanz F, Furlong LI. The DisGeNET cytoscape app: exploring and visualizing disease genomics data. Computational and Structural Biotechnology Journal. 2021;19:2960–2967.

[79] Hunter JD. Matplotlib: a 2D graphics environment. Computing in Science & Engineering. 2007;9(3):90–95.

[80] Waskom ML. Seaborn: statistical data visualization. Journal of Open Source Software. 2021;6(60):3021. Available from: https://doi.org/10.21105/joss.03021.

[81] McKinney W, et al. Data structures for statistical computing in Python. In: Proceedings of the 9th Python in science conference. vol. 445. Austin, TX; 2010. p. 51–56.

[82] Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585:357–362.

[83] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods. 2020;17:261–272.

[84] Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using NetworkX. In: Varoquaux G, Vaught T, Millman J, editors. Proceedings of the 7th Python in Science conference. Pasadena, CA USA; 2008. p. 11 – 15.

[85] Anders S, Huber W. Differential expression analysis for sequence count data. Nature Precedings. 2010:1–1.

[86] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15(12):1–21.

[87] Morris JH, Lotia S, Wu A, Doncheva NT, Albrecht M, Pico AR, et al. Setsapp for Cytoscape: set operations for Cytoscape nodes and edges. F1000Research. 2014;3:149.

[88] Beaulieu JM, Espinoza S, Gainetdinov RR. Dopamine receptors – IUPHAR review 13. British Journal of Pharmacology. 2015;172(1):1–23.

[89] Zhao L, Lin Y, Lao G, Wang Y, Guan L, Wei J, et al. Association study of dopamine receptor genes polymorphism with cognitive functions in bipolar I disorder patients. Journal of Affective Disorders. 2015;170:85–90.

[90] Maréchal A, Zou L. RPA-coated single-stranded DNA as a platform for post-translational modifications in the DNA damage response. Cell Research. 2015;25(1):9–23.

[91] Gupta A, Schulze TG, Nagarajan V, Akula N, Corona W, Jiang X, et al. Interaction networks of lithium and valproate molecular targets reveal a striking enrichment of apoptosis functional clusters and neurotrophin signaling. The Pharmacogenomics Journal. 2012;12(4):328–341.

[92] Weichenrieder O, Wild K, Strub K, Cusack S. Structure and assembly of the Alu domain of the mammalian signal recognition particle. Nature. 2000;408(6809):167–173.

[93] Faoro C, Ataide SF. Noncanonical functions and cellular dynamics of the mammalian signal recognition particle components. Frontiers in Molecular Biosciences. 2021;8:420.

[94] Smidt MP, Smits SM, Burbach JPH. Homeobox gene Pitx3 and its role in the development of dopamine neurons of the substantia nigra. Cell and Tissue Research. 2004;318(1):35–43.

[95] Smidt MP, Burbach JPH. How to make a mesodiencephalic dopaminergic neuron. Nature Reviews Neuroscience. 2007;8(1):21–32.

[96] Smidt MP, Van Schaick HS, Lanctôt C, Tremblay JJ, Cox JJ, Van Der Kleij AA, et al. A homeodomain gene Ptx3 has highly restricted brain expression in mesencephalic dopaminergic neurons. Proceedings of the National Academy of Sciences. 1997;94(24):13305–13310.

[97] Xu W, Cohen-Woods S, Chen Q, Noor A, Knight J, Hosang G, et al. Genome-wide association study of bipolar disorder in Canadian and UK populations corroborates disease loci including SYNE1 and CSMD1. BMC Medical Genetics. 2014;15(1):1–13.

[98] Dowhan DH, Hong EP, Auboeuf D, Dennis AP, Wilson MM, Berget SM, et al. Steroid hormone receptor coactivation and alternative RNA splicing by U2AF65-related proteins CAPERα and CAPERβ. Molecular Cell. 2005;17(3):429–439.

[99] Li YI, Van De Geijn B, Raj A, Knowles DA, Petti AA, Golan D, et al. RNA splicing is a primary link between genetic variation and disease. Science. 2016;352(6285):600–604.

[100] Ng MM, Dippold HC, Buschman MD, Noakes CJ, Field SJ. GOLPH3L antagonizes GOLPH3 to determine Golgi morphology. Molecular Biology of the Cell. 2013;24(6):796–808.

[101] Hauberg ME, Fullard JF, Zhu L, Cohain AT, Giambartolomei C, Misir R, et al. Differential activity of transcribed enhancers in the prefrontal cortex of 537 cases with schizophrenia and controls. Molecular Psychiatry. 2019;24(11):1685–1695.

[102] Cross-Disorder Group of the Psychiatric Genomics Consortium and others. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nature Genetics. 2013;45(9):984.

[103] Ciarimboli G, Struwe K, Arndt P, Gorboulev V, Koepsell H, Schlatter E, et al. Regulation of the human organic cation transporter hOCT1. Journal of Cellular Physiology. 2004;201(3):420–428.

[104] Koepsell H. General overview of organic cation transporters in brain. In: Daws LC, editor. Organic cation transporters in the central nervous system. Cham: Springer International Publishing; 2021. p. 1–39. Available from: https://doi.org/10.1007/164_2021_449.

[105] De Vega S, Iwamoto T, Yamada Y. Fibulins: multiple roles in matrix structures and tissue functions. Cellular and Molecular Life Sciences. 2009;66(11):1890–1902.

[106] Radice PD, Mathieu P, Leal MC, Farías MI, Ferrari C, Puntel M, et al. Fibulin-2 is a key mediator of the pro-neurogenic effect of TGF-beta1 on adult neural stem cells. Molecular and Cellular Neuroscience. 2015;67:75–83.

[107] Silver NC, Dunlap WP. Averaging correlation coefficients: should Fisher's z transformation be used? Journal of Applied Psychology. 1987;72(1):146–148.

[108] Strube MJ. Averaging correlation coefficients: influence of heterogeneity and set size. Journal of Applied Psychology. 1988;73(3):559.

[109] Morselli Gysi D, de Miranda Fragoso T, Zebardast F, Bertoli W, Busskamp V, Almaas E, et al. Whole transcriptomic network analysis using Co-expression Differential Network Analysis (CoDiNA). PLOS ONE. 2020;15(10):e0240523.

[110] Field AP. Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed-and random-effects methods. Psychological Methods. 2001;6(2):161.

[111] Gulla M. An integrated systems biology approach to investigate transcriptomic data of thyroid carcinoma. Norges teknisk-naturvitenskapelige universitet (NTNU); 2019. Master's thesis available from: http://hdl.handle.net/11250/2621725.

[112] Baralle FE, Giudice J. Alternative splicing as a regulator of development and tissue identity. Nature Reviews Molecular Cell Biology. 2017;18(7):437–451.

[113] Hung MC, Link W. Protein localization in disease and therapy. Journal of Cell Science. 2011;124(20):3381–3392.

[114] Mehler MF, Gokhan S. Developmental mechanisms in the pathogenesis of neurodegenerative diseases. Progress in Neurobiology. 2001;63(3):337–363.

[115] Chen H, DeLong C, Bame M, Rajapakse I, Herron T, McInnis M, et al. Transcripts involved in calcium signaling and telencephalic neuronal fate are altered in induced pluripotent stem cells from bipolar disorder patients. Translational Psychiatry. 2014;4(3):e375–e375.

[116] Madison JM, Zhou F, Nigam A, Hussain A, Barker DD, Nehme R, et al. Characterization of bipolar disorder patient-specific induced pluripotent stem cells from a family reveals neurodevelopmental and mRNA expression abnormalities. Molecular Psychiatry. 2015;20(6):703–717.

[117] Liszewska E, Jaworski J. Neural stem cell dysfunction in human brain disorders. Human Neural Stem Cells. 2018:283–305.

[118] Tkachev D, Mimmack ML, Ryan MM, Wayland M, Freeman T, Jones PB, et al. Oligodendrocyte dysfunction in schizophrenia and bipolar disorder. The Lancet. 2003;362(9386):798–805.

[119] Konopaske GT, Lange N, Coyle JT, Benes FM. Prefrontal cortical dendritic spine pathology in schizophrenia and bipolar disorder. JAMA Psychiatry. 2014;71(12):1323–1331.

[120] Deckersbach T, Dougherty DD, Savage C, McMurrich S, Fischman AJ, Nierenberg A, et al. Impaired recruitment of the dorsolateral prefrontal cortex and hippocampus during encoding in bipolar disorder. Biological Psychiatry. 2006;59(2):138–146.

[121] Nakatani N, Hattori E, Ohnishi T, Dean B, Iwayama Y, Matsumoto I, et al. Genome-wide expression analysis detects eight genes with robust alterations specific to bipolar I disorder: relevance to neuronal network perturbation. Human Molecular Genetics. 2006;15(12):1949–1962.

[122] Sun X, Wang JF, Tseng M, Young LT. Downregulation in components of the mitochondrial electron transport chain in the postmortem frontal cortex of subjects with bipolar disorder. Journal of Psychiatry and Neuroscience. 2006;31(3):189–196.

[123] Naumova OY, Lee M, Rychkov SY, Vlasova NV, Grigorenko EL. Gene expression in the human brain: the current state of the study of specificity and spatiotemporal dynamics. Child Development. 2013;84(1):76–88.

[124] Simillion C, Liechti R, Lischer HE, Ioannidis V, Bruggmann R. Avoiding the pitfalls of gene set enrichment analysis with SetRank. BMC Bioinformatics. 2017;18(1):1–14.

[125] Dalmer TR, Clugston RD. Gene ontology enrichment analysis of congenital diaphragmatic hernia-associated genes. Pediatric Research. 2019;85(1):13–19.

[126] Héroux ME. Analyzing dependent data as if independent biases effect size estimates and increases the risk of false-positive findings. Journal of Applied Physiology. 2021;130(3):675–676.

[127] Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nature Reviews Genetics. 2018;19(5):299–310.

[128] Akiyama M. Multi-omics study for interpretation of genome-wide association study. Journal of Human Genetics. 2021;66(1):3–10.

Appendix A

Supplement to Bipolar Disorder and Control Data Sets

A.1 Normalization of Microarrays and RNA-Seq

The construction of a consensus CSD network for BP relative to control samples was based on six individual studies. The normalization methods used in these studies are summarized in Table A.1.

Table A.1: Normalization methods for bipolar disorder data sets from Gene Expression Omnibus (GEO). All studies contained samples originating from the dorsolateral prefrontal cortex. RMA: Robust Multi-array Analysis, MAS5: Micro Array Suite 5.0

GEO accession	Platform	Normalization in study	Additional normalization	Reference
GSE80655	Illumina HiSeq 2000	No information about normalization	Between-sample normalization with DESeq [85, 86]	[68]
GSE92538	Affymetrix HG-U133A or HG-U133 Plus 2	Log(2)-transformed RMA and quantile normalized, gender-checked and per-batch median-centered.	None	[69]
GSE53987	Affymetrix HG-U133 Plus 2	RMA normalized.	None	[70]
GSE5388	Affymetrix HG-U133A	RMA normalized.	None	[71]
GSE12649	Affymetrix HG-U133A	Processed by MAS5 and normalized by median centering.	None	[72]
GSE120340	Affymetrix HG-U133 Plus 2	Log(2)-transformed, RMA normalized.	None	[73]

A.2 Patient Information

Several of the studies used to generate a consensus CSD network for BP received brain tissue from the same brain bank. This could mean that these studies are based on the same patients. A mapping of age, gender and age of onset for the included BP and control samples from GSE12649 [72] and GSE5388 [71] is included in Table A.2 and A.3, respectively. GSE120340 [73] has received brain tissue from the the same brain bank as GSE12649 [72] and GSE5388 [71], but is omitted from the patient mapping due to lack of available patient information. A mapping of age, gender and ethnicity for the included BP and control samples from GSE80655 [68] and GSE92538 [69] is included in Table A.4 and A.5, respectively. No patient information from GSE53967 [70] is included here as it is the only one of the included studies that has received brain tissue from the University of Pittsburgh. Additional information about the patients, such as cause of death, brain weight, medication, drug abuse, alcohol abuse and PMI, are in some cases available form the studies and interested readers are referred to the relevant references for more information [68–73].

Table A.2: Patient information mapping for bipolar disorder samples from GSE12649 [72] and GSE5388 [71], which both have received samples from Stanley Medical Research Institute. Patients with same age, gender and age of onset in the two studies are indicated in green. Patients that must be unique to one study are indicated in red.

GSE12649			GSE5388		
Age	**Gender**	**Age of onset**	**Age**	**Gender**	**Age of onset**
19	Male	17	19	Male	17
29	Male	17	29	Male	17
29	Male	22	29	Male	22
29	Female	18	29	Female	18
33	Female	15	33	Female	15
35	Female	21	35	Female	21
35	Male	19	35	Male	19
35	Male	14			
41	Male	21	41	Male	21
			41	Male	22
41	Female	14	41	Female	14
42	Male	18	42	Male	18
42	Female	20	42	Female	20
43	Female	25	43	Female	25
			43	Female	29
44	Male	33	44	Male	33
44	Female	26	44	Female	26
45	Male	16	45	Male	16
45	Male	35	45	Male	35
48	Female	33	48	Female	33
48	Male	31	48	Male	31
49	Female	22			
49	Female	20	49	Female	20
50	Female	25	50	Female	25
51	Female	35			
51	Male	23	51	Male	23
54	Male	45	54	Male	45
55	Female	40			
56	Male	28	56	Male	28
56	Female	14	56	Female	14
58	Female	27	58	Female	27
59	Male	25	59	Male	25
59	Female	48			
63	Female	43	63	Female	43
64	Male	19	64	Male	19

Table A.3: Patient information mapping for control samples from GSE12649 [72] and GSE5388 [71], which both have received samples from Stanley Medical Research Institute. Patients with same age and gender in the two studies are indicated in green or in yellow if there are multiple potential matches. Patients that must be unique to one study are indicated in red.

GSE12649		GSE5388	
Age	**Gender**	**Age**	**Gender**
31	Male	31	Male
32	Male	32	Male
		32	Male
33	Female		
34	Male	34	Male
34	Female	34	Female
35	Male	35	Male
35	Male		
37	Male	37	Male
38	Female	38	Female
38	Female	38	Female
39	Female	39	Female
40	Male	40	Male
41	Female		
42	Male	42	Male
44	Female	44	Female
44	Female	44	Female
45	Male	45	Male
45	Male	45	Male
46	Male	46	Male
47	Male	47	Male
47	Male	47	Male
		47	Male
48	Male	48	Male
48	Male	48	Male
		48	Male
49	Male	49	Male
49	Male		
49	Female	49	Female
50	Male	50	Male
51	Male	51	Male
51	Male		
		52	Male
53	Male	53	Male
53	Male		
55	Male	55	Male
57	Male		
		59	Male
60	Male		

Table A.4: Patient information mapping for bipolar disorder samples from GSE80655 [68] and GSE92538 [69], which both have received samples from Pritzker Neuropsychiatric Disorders Research Consortium. Patients with same age, gender and ethnicity in the two studies are indicated in green or in yellow if there are multiple potential matches. Patients that must be unique to one study are indicated in red.

GSE80655			GSE92538		
Age	**Gender**	**Ethnicity**	**Age**	**Gender**	**Ethnicity**
23	Male	Caucasian	23	Male	Caucasian
25	Female	Caucasian			
26	Male	Caucasian	26	Male	Caucasian
32	Male	Caucasian	32	Male	Caucasian
			33	Male	Caucasian
			34	Female	Caucasian
36	Female	Caucasian	36	Female	Caucasian
			36	Male	Caucasian
			39	Female	Caucasian
40	Male	Caucasian	40	Male	Caucasian
42	Male	Caucasian			
49	Male	Other	49	Male	Other
49	Female	Caucasian	49	Female	Caucasian
51	Female	Caucasian	51	Female	Caucasian
51	Female	Caucasian			
51	Male	Caucasian			
52	Male	Caucasian	52	Male	Caucasian
52	Male	Caucasian			
52	Male	Caucasian			
53	Male	Caucasian			
56	Female	Caucasian	56	Female	Caucasian
59	Female	Caucasian	59	Female	Caucasian
59	Male	Caucasian	59	Male	Caucasian
			63	Female	Caucasian
63	Male	Caucasian	63	Male	Caucasian
			66	Female	Caucasian
			68	Female	Caucasian
69	Male	Caucasian	69	Male	Caucasian
			69	Male	Caucasian
69	Female	Caucasian	69	Female	Caucasian
70	Male	Caucasian	70	Male	Caucasian
			73	Male	Caucasian
			81	Female	Caucasian

Table A.5: Patient information mapping for control samples from GSE80655 [68] and GSE92538 [69], which both have received samples from Pritzker Neuropsychiatric Disorders Research Consortium. Patients with same age, gender and ethnicity in the two studies are indicated in green or in yellow if there are multiple potential matches. Patients that must be unique to one study are indicated in red.

GSE80655			GSE92538		
Age	**Gender**	**Ethnicity**	**Age**	**Gender**	**Ethnicity**
			18	Male	Caucasian
			19	Male	Caucasian
			25	Male	Caucasian
			30	Male	Caucasian
32	Male	Caucasian	32	Male	Caucasian
32	Male	Caucasian	32	Male	Caucasian
35	Male	Caucasian	35	Male	Caucasian
39	Male	Pacific Islander			
39	Male	Caucasian	39	Male	Caucasian
40	Male	Caucasian	40	Male	Caucasian
40	Male	Caucasian	40	Male	Caucasian
41	Male	Caucasian	41	Male	Caucasian
43	Male	Asian	43	Male	Asian
44	Male	Caucasian	44	Male	Caucasian
45	Male	Caucasian	45	Male	Caucasian
45	Female	Caucasian	45	Female	Caucasian
			47	Female	Asian
			47	Female	Caucasian
48	Male	Caucasian	48	Male	Caucasian
			48	Male	Caucasian
			48	Female	Caucasian
49	Male	Caucasian	49	Male	Caucasian
			50	Male	Caucasian
			50	Male	Asian
			52	Male	Caucasian
			52	Male	Caucasian
			52	Female	Caucasian
			53	Male	Caucasian
			54	Male	Caucasian
			54	Male	Caucasian
55	Male	Caucasian	55	Male	Caucasian
			55	Male	Caucasian
			55	Male	Caucasian
			55	Male	Caucasian
56	Male	Caucasian	56	Male	Caucasian
			56	Male	Caucasian
			56	Male	Caucasian
			57	Female	Caucasian
58	Male	Caucasian	58	Male	Caucasian
			58	Male	Caucasian
			58	Male	Caucasian
			59	Male	African American
			59	Male	Caucasian
			60	Male	Caucasian
			60	Female	Caucasian
			60	Male	Caucasian
			62	Female	Caucasian

Continued on next page

Table A.5 – continued from previous page

GSE80655			GSE92538		
Age	**Gender**	**Ethnicity**	**Age**	**Gender**	**Ethnicity**
			62	Female	Caucasian
63	Male	Caucasian	63	Male	Caucasian
			63	Male	Caucasian
			63	Male	Caucasian
			63	Female	Caucasian
			64	Male	Caucasian
			64	Male	Caucasian
64	Female	Caucasian	64	Female	Caucasian
			64	Female	Caucasian
			64	Male	Caucasian
			65	Female	Caucasian
65	Male	African American	65	Male	African American
66	Male	Caucasian	66	Male	Caucasian
67	Male	Caucasian	67	Male	Caucasian
			67	Male	Caucasian
			67	Male	Asian
			68	Female	Caucasian
			68	Female	Caucasian
			69	Male	Caucasian
			69	Male	Caucasian
70	Male	Caucasian	70	Male	Caucasian
70	Female	Caucasian	70	Female	Caucasian
			70	Female	Caucasian
			70	Female	Caucasian
			71	Male	Caucasian
			71	Male	Caucasian
			72	Female	Caucasian
			72	Male	Caucasian
			73	Female	Caucasian
			73	Female	Caucasian
			73	Female	Caucasian
			75	Male	Caucasian
			77	Male	Caucasian
			77	Male	Caucasian
			78	Male	Caucasian
			78	Female	Caucasian
			79	Male	Caucasian

Appendix B

Overview of Software Versions

The development, testing, network creation and analysis in this thesis have relied on several third-party software resources. A complete list with version numbers is provided in Table B.1. Analyses relying on the steps from the CSD approach [9] have used older versions of the software.

Table B.1: List of software version numbers used in analyses. Libraries/apps have been grouped with their associated programming language/software platform. Note that scripts from the CSD approach [9] relied on older versions of the software (not listed).

Software	Version
Python	3.9.7
Matplotlib	3.4.3
Seaborn	0.11.2
Matplotlib-ven	0.11.6
Pandas	1.3.4
Numpy	1.20.3
SciPy	1.7.1
NetworkX	2.7
R	3.6.1
DESeq	1.26.0
Excel	2202 (Build 16.0.14931.20128)
Cytoscape	3.7.2
DisGeNet-app	7.3.0
SetsApp	2.2.0
PANTHER	17.0

Appendix C

Supplement to Method Development

C.1 Pairwise Comparison of Correlation of Correlations

The method development in this thesis has been devoted to construction of consensus CSD networks and has focused on combination of correlation coefficients using either weighted untransformed or Fisher's Z transformed averages. Spearman rank correlation coefficients were calculated for these combined correlation coefficients relative to the correlation coefficients from the reference data set (a 1000-gene data set originating from "*Skin - Not Sun Exposed (Suprapubic)*"). A Wilcoxon signed-rank test indicated that these Spearman rank correlations were significantly different for the two combination methods. A plot of the pairwise differences and the distribution of the differences are displayed in Figure C.1. Note that the distribution of the differences is approximately symmetrical, thus justifying the use of a Wilcoxon signed-rank test.

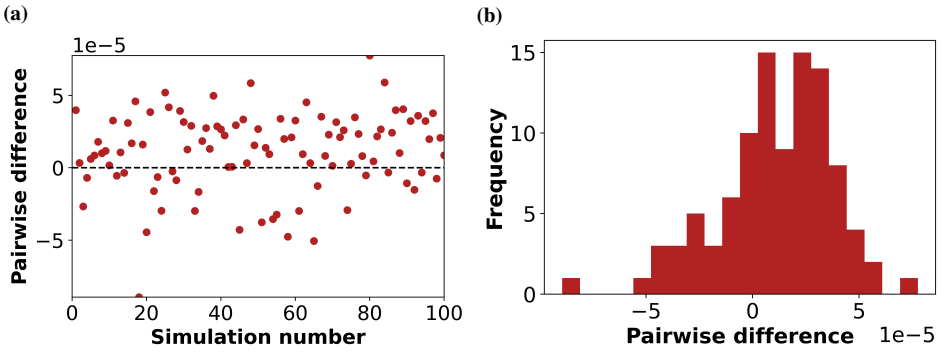


Figure C.1: (a) Scatter plot and (b) histogram of pairwise differences between correlation of correlations for weighted untransformed and Fisher's Z transformed averages. Correlation of correlations has been calculated as the Spearman rank correlation coefficient of a combination method (weighted untransformed or Fisher's Z transformed averages) relative to reference correlations from a 1000-gene data set originating from "*Skin - Not Sun Exposed (Suprapubic)*".

C.2 Pairwise Comparison of RMSEs

The methods for combining correlation coefficients (Fisher's Z transformed or weighted untransformed averages) have been evaluated by calculation of RMSEs relative to the correlation coefficients from the reference data set (a 1000-gene data set originating from "*Skin - Not Sun Exposed (Suprapubic)*"). A Wilcoxon signed-rank test indicated that these RMSEs were significantly different for the two combination methods. A plot of the pairwise differences and the distribution of the differences between the RMSEs are displayed in Figure C.2. The distribution of the differences is nearly symmetrical, thus justifying the use of a Wilcoxon signed-rank test.

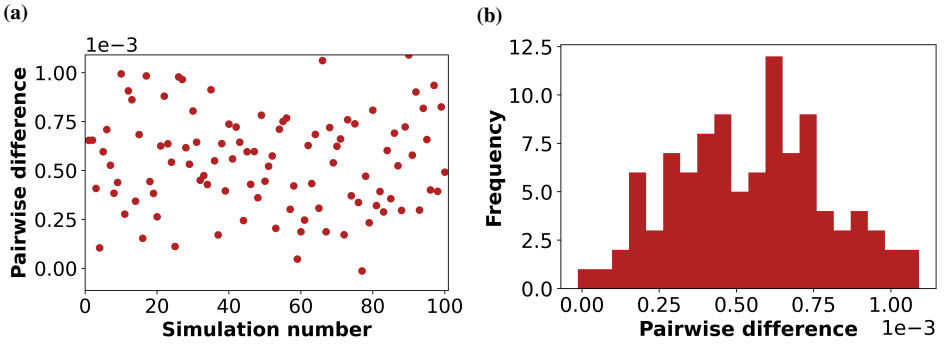


Figure C.2: (a) Scatter plot and (b) histogram of pairwise differences between root mean square errors (RMSEs) for weighted untransformed and Fisher's Z transformed averages of correlation coefficients. The RMSEs were calculated relative to the reference correlations from a 1000-gene data set originating from "*Skin - Not Sun Exposed (Suprapubic)*".

C.3 Jaccard Indices for Combined Correlation Coefficients

The overlap between the top n most strongly co-expressed gene pairs based on the combined correlation coefficients from 1000-gene data sets originating from "*Skin - Not Sun Exposed (Suprapubic)*" has been evaluated by calculating Jaccard indices. The Jaccard index for the weighted untransformed averages relative to Fisher's Z transformed averages of correlation coefficients is plotted as a function of number of investigated gene pairs in Figure C.3.

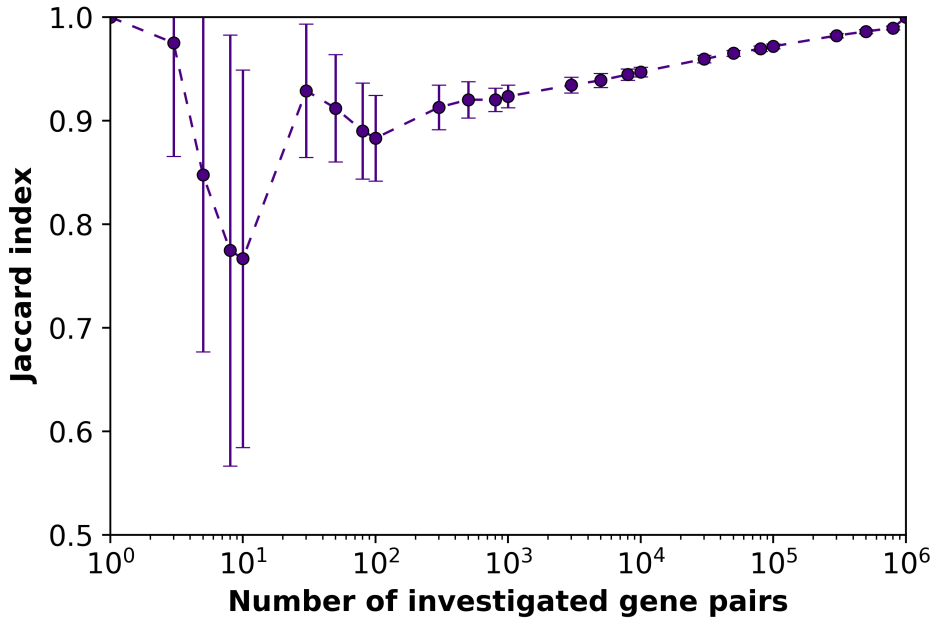


Figure C.3: Jaccard index as a function of number of investigated gene pairs for combined Spearman rank correlation coefficients based on weighted untransformed averages relative to Fisher's Z transformed averages of correlation coefficients from 1000-gene data sets originating from "*Skin - Not Sun Exposed (Suprapubic)*". Error bars represent standard deviations.

Appendix D

Clustering Analysis of Control Samples

The construction of consensus CSD networks for BP relied on the combination of six data sets. As some of these sets have received tissue from the same brain bank, a clustering analysis was conducted. The hierarchically-clustered heat map for the control data sets is shown in Figure D.1.

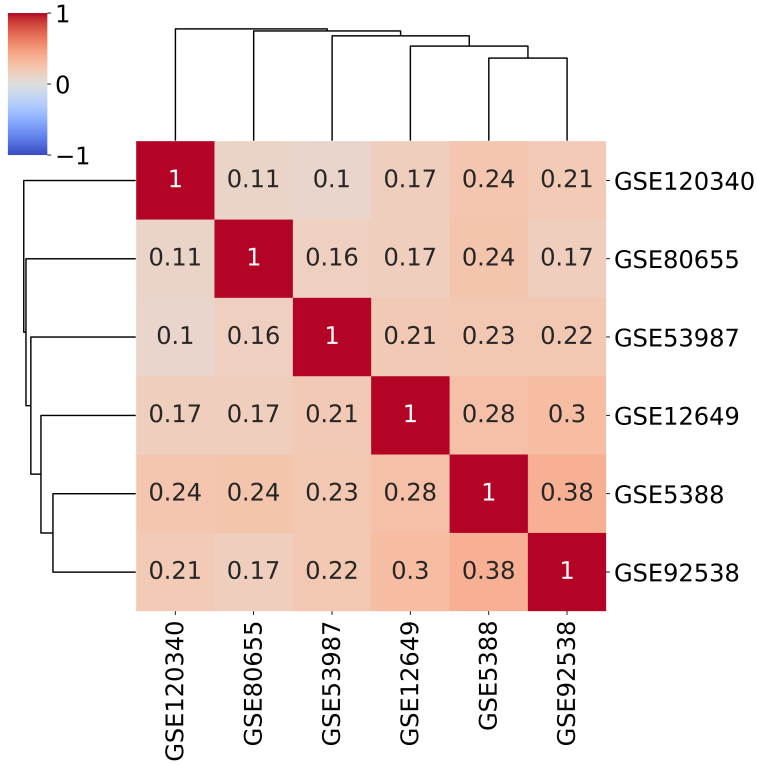


Figure D.1: Hierarchically-clustered heat map for control data sets. The elements of the matrix correspond to the pairwise Spearman rank correlation coefficients between the Spearman rank correlation coefficients from the two indicated studies.

Appendix E

Supplement to Model Comparison

E.1 Model Comparison at the Level of Correlations – Control Samples

The construction of a consensus CSD network for BP involved calculation of combined correlation coefficients. The combined correlation coefficients can be estimated by the use of either weighted untransformed or Fisher's Z transformed averages. The resulting estimates from the combination methods have been plotted against each other in Figure E.1 for control samples.

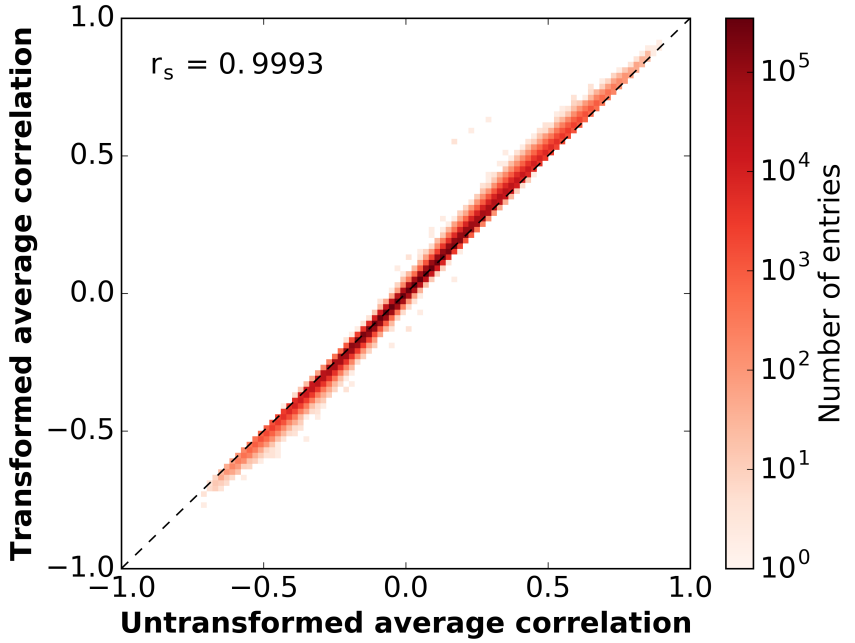
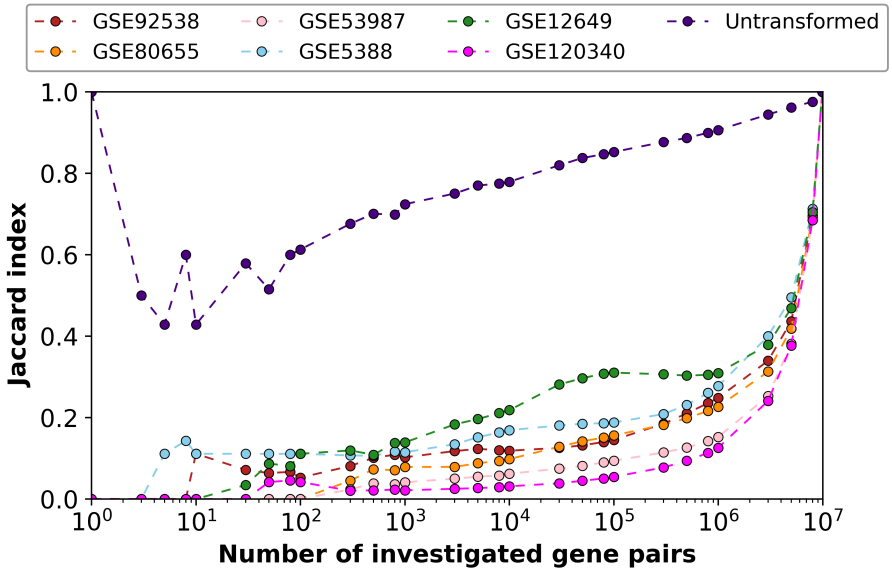


Figure E.1: Heat map between combined Spearman rank correlation coefficients based on Fisher's Z transformed and weighted untransformed averages for control samples. The dashed line represents the expected relationship ($y = x$). r_s represents the Spearman rank correlation coefficient between the combined correlations.

E.2 Jaccard Indices for Correlation Coefficients in Bipolar Disorder and Control Samples

The combined correlation coefficients from BP and control samples have been compared to each other and to underlying data sets by calculation of Jaccard indices of top n most strongly co-expressed gene pairs. This is displayed in Figure E.2 using Fisher's Z transformed averages as reference values.

(a)



(b)

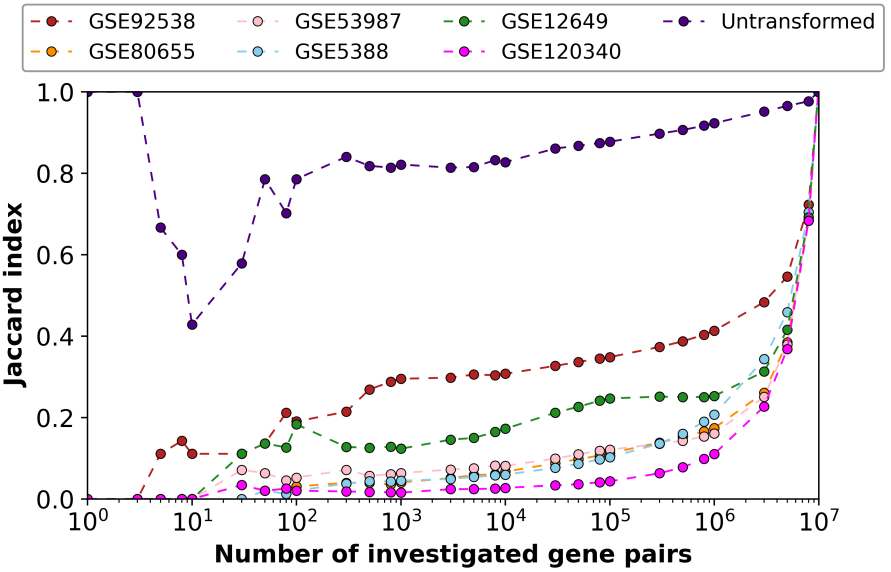


Figure E.2: Jaccard index as a function of number of investigated gene pairs between Spearman rank correlation coefficients from indicated data sets and Fisher's Z transformed averages originating from (a) bipolar disorder and (b) control samples. The term "Untransformed" refers to weighted untransformed averages of the correlation coefficients.

E.3 Similarity of Neighbourhoods in the CSD Networks for Bipolar Disorder

The construction of consensus CSD networks for BP has been based on two different combination methods: weighted untransformed and Fisher's Z transformed averages of correlation coefficients. The similarity between the neighbourhoods in the resulting CSD networks is presented in Figure E.3 using the network based on Fisher's Z transformed averages of correlation coefficients as reference.

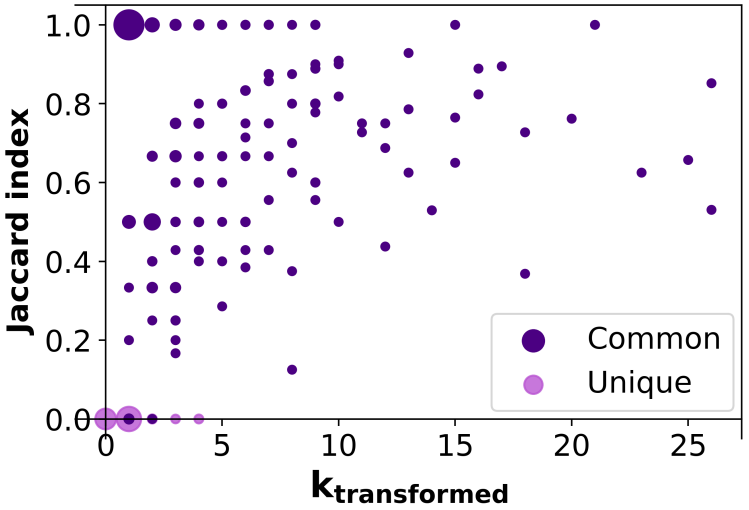


Figure E.3: Comparison of neighbourhoods in the CSD networks for bipolar disorder using Fisher's Z transformed averages of correlation coefficients as reference. The nodes are organized by degree in the reference network (denoted by $k_{transformed}$) and Jaccard indices are calculated between the neighbours of node i in this networks relative to the CSD network based on weighted untransformed averages. The terms "common" and "unique" refer to whether the nodes are shared between the two CSD networks or are unique to the reference network. The size of the points in the plot reflects the number of nodes in the reference network with the given characteristic.

Appendix F

Structural Analysis of CSD Network Based on Fisher's Z Transformed Averages

Structural analyses of the CSD networks for BP have been conducted to evaluate if these networks are similar to typical CSD networks. The CSD network based on weighted untransformed averages of correlation coefficients has been the main focus in this thesis. However, structural analyses of the CSD network based on Fisher's Z transformed averages are presented here. This includes analyses of degree distribution (Figure F.1), assortativity (Table F.1), average clustering coefficients (Table F.1) and node homogeneity scores (Figure F.2).

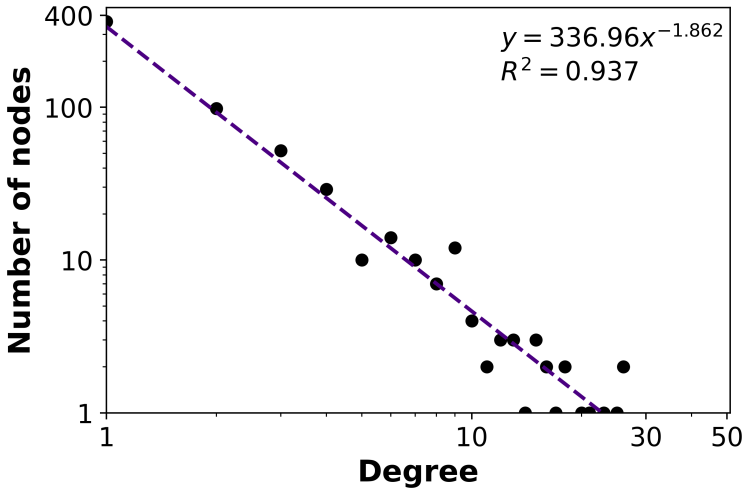


Figure F.1: Degree distribution of the consensus CSD network for bipolar disorder based on Fisher's Z transformed averages of correlation coefficients. The purple, dashed line represents the fitted power law.

Table F.1: Assortativity and average clustering in the CSD network for bipolar disorder based on Fisher's Z transformed averages of correlation coefficients.

Network type	Degree assortativity coefficient	Average clustering coefficient
Full CSD network	0.097	0.088
C network	-0.090	0.40
S network	-0.17	0.025
D network	-0.15	0

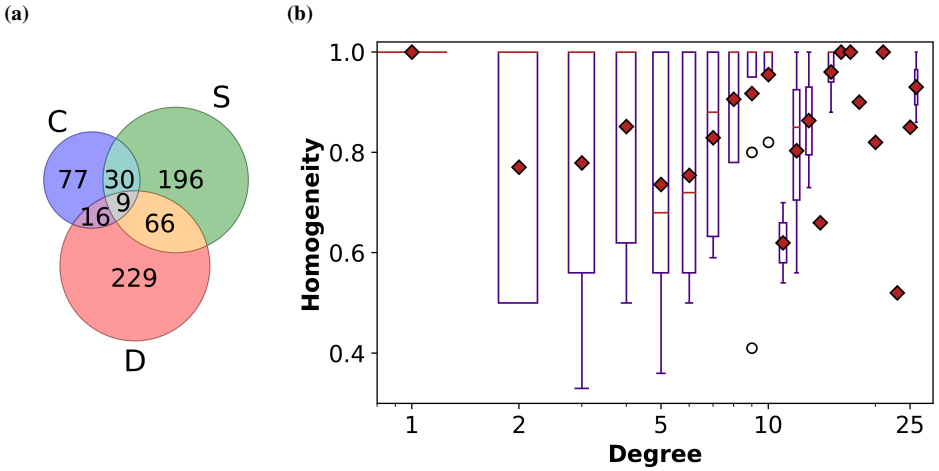


Figure F.2: (a) Number of nodes involved in each type of interaction and (b) node homogeneity scores in the CSD network for bipolar disorder based on Fisher's Z transformed averages of correlation coefficients. Red bars and red diamonds indicate median and mean homogeneity scores, respectively.

Appendix G

Supplement to Functional Analysis

G.1 Disease Enrichment

The CSD network for BP was subjected to disease enrichment for BP, which indicated that 29 genes in the CSD network had previously been associated with BP. These genes, as well as their degrees and homogeneity scores, are provided in Table G.1.

Table G.1: Genes, and their degrees and homogeneity scores, from disease enrichment of the CSD network for bipolar disorder. The rows are coloured according to the main interaction type of the nodes, where C, S and D types are coloured blue, green and red, respectively. Nodes with equal proportions of two interaction types are not coloured.

Node	k	k_C	k_S	k_D	H
AGT	18	18	0	0	1.0
MLC1	15	14	1	0	0.88
S1PR1	12	11	1	0	0.85
NR2E1	9	8	1	0	0.8
DRD4	6	0	1	5	0.72
GLO1	5	0	5	0	1.0
ALDH2	4	4	0	0	1.0
SST	4	1	3	0	0.62
ADRA2C	2	0	1	1	0.5
ATP1A1	2	1	1	0	0.5
CSRP1	2	2	0	0	1.0
DPYSL2	2	0	1	1	0.5
GRM3	2	0	0	2	1.0
VGF	2	0	2	0	1.0
ADM	1	0	1	0	1.0
APOD	1	0	1	0	1.0
ITIH1	1	0	1	0	1.0
CNR2	1	0	0	1	1.0
CNTN6	1	0	1	0	1.0
GNB3	1	0	1	0	1.0
PC	1	0	1	0	1.0
GPRC5D	1	0	1	0	1.0
IL6	1	0	1	0	1.0
PVALB	1	0	1	0	1.0
TAC1	1	0	1	0	1.0
HPGDS	1	0	0	1	1.0
IL2RB	1	0	0	1	1.0
SPR	1	0	0	1	1.0
RELN	1	0	1	0	1.0

G.2 GO Enrichment of Communities

Communities in the CSD network for BP were identified using the Louvain algorithm [29]. The communities were then analysed for overrepresentation of GO terms. Three of the communities (number 5, 8 and 13 in Figure 4.13) were enriched for biological processes when requiring FDR < 0.05. One additional community (number 10 in Figure 4.13) were enriched for biological processes when using FDR < 0.1. Table G.2, G.3, G.4 and G.5 provide complete lists of all enriched GO terms for the four communities.

Table G.2: Complete Gene Ontology (GO) enrichment of community number 5 in the CSD network for bipolar disorder. Only results with false discovery rate (FDR) < 0.05 are displayed.

GO biological process	**Number of genes**	**Fold enrichment**	**Raw *P* value**	**FDR**
Proton transmembrane transport (GO:1902600)	6	15.51	4.70E-06	2.53E-02
ATP metabolic process (GO:0046034)	7	11.63	3.82E-06	4.11E-02

Table G.3: Complete Gene Ontology (GO) enrichment of community number 8 in the CSD network for bipolar disorder. Note that related classes of an ontology are grouped together. Only results with false discovery rate (FDR) < 0.05 are displayed.

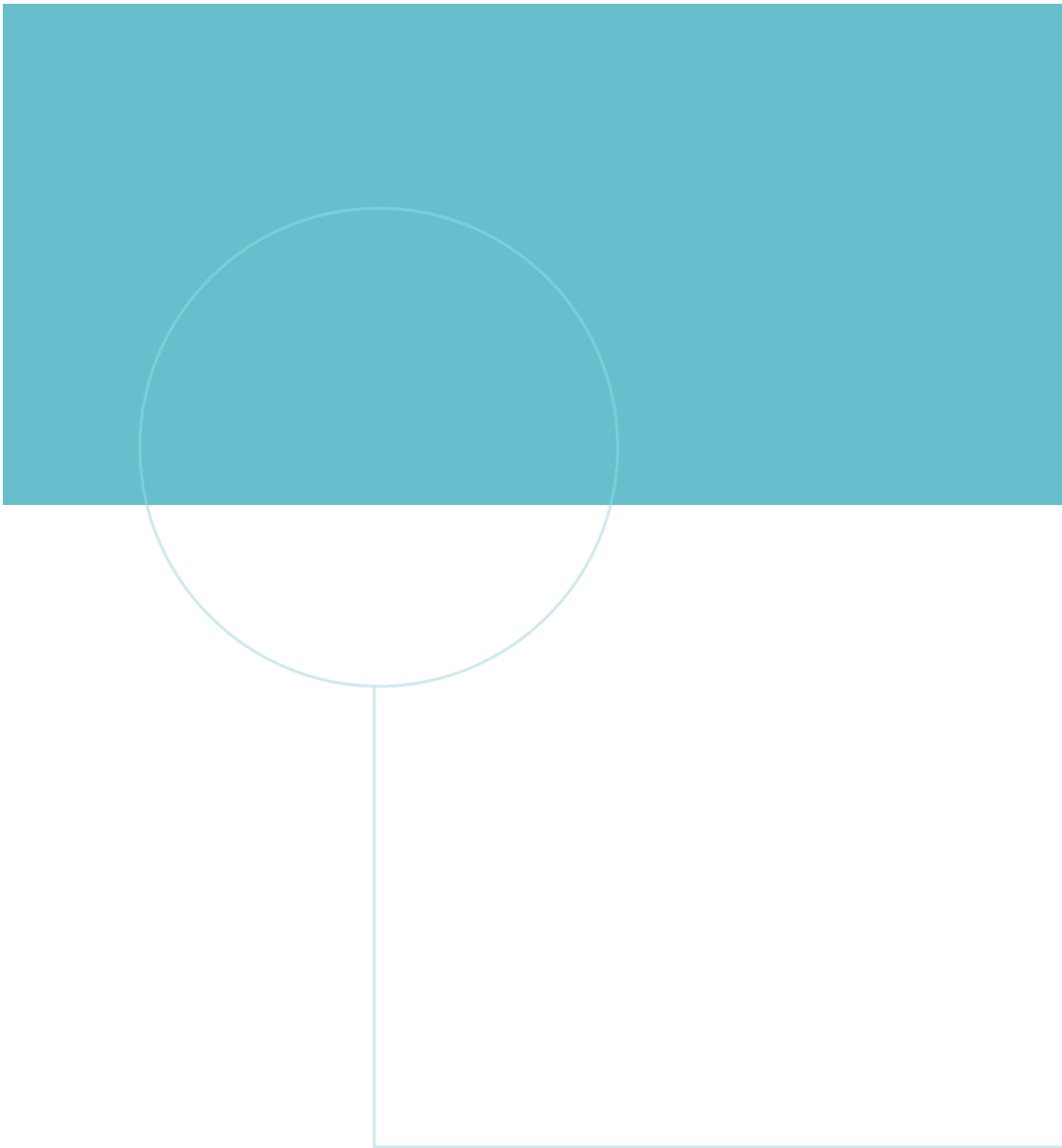
GO biological process	Number of genes	Fold enrichment	Raw *P* value	FDR
Vertebrate eye-specific patterning (GO:0150064)	2	> 100	3.02E-05	2.71E-02
Neuron remodeling (GO:0016322)	2	> 100	4.53E-05	3.48E-02
Neuron maturation (GO:0042551)	2	> 100	1.09E-04	4.87E-02
Complement-mediated synapse pruning (GO:0150062)	2	> 100	4.53E-05	3.25E-02
Synapse pruning (GO:0098883)	3	> 100	4.61E-07	4.96E-03
Cell junction disassembly (GO:0150146)	3	> 100	6.33E-07	3.41E-03
Cell junction organization (GO:0034330)	4	24.34	9.00E-06	1.38E-02
Synapse organization (GO:0050808)	3	33.40	7.31E-05	3.75E-02
Complement activation, classical pathway (GO:0006958)	3	> 100	3.11E-06	1.12E-02
Humoral immune response mediated by circulating immunoglobulin (GO:0002455)	3	> 100	3.11E-06	8.38E-03
Immune response (GO:0006955)	5	8.25	6.00E-05	3.80E-02
Immunoglobulin mediated immune response (GO:0016064)	3	49.06	2.47E-05	2.66E-02
B cell mediated immunity (GO:0019724)	3	49.06	2.47E-05	2.41E-02
Adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains (GO:0002460)	3	34.89	6.46E-05	3.48E-02
Lymphocyte mediated immunity (GO:0002449)	3	35.68	6.06E-05	3.62E-02
Immune effector process (GO:0002252)	4	19.38	2.16E-05	2.59E-02
Complement activation (GO:0006956)	3	65.42	1.11E-05	1.49E-02
Activation of immune response (GO:0002253)	4	31.24	3.46E-06	7.44E-03
Positive regulation of immune response (GO:0050778)	4	16.23	4.29E-05	3.55E-02
Regulation of immune system process (GO:0002682)	5	8.15	6.38E-05	3.61E-02
Positive regulation of immune system process (GO:0002684)	5	12.28	8.66E-06	1.55E-02
Innate immune response (GO:0045087)	4	13.86	7.87E-05	3.69E-02
Response to biotic stimulus (GO:0009607)	5	8.31	5.82E-05	3.92E-02
Defense response (GO:0006952)	5	7.88	7.51E-05	3.68E-02

Table G.4: Complete Gene Ontology (GO) enrichment of community number 10 in the CSD network for bipolar disorder. Note that related classes of an ontology are grouped together. Only results with false discovery rate (FDR) < 0.1 are displayed.

GO biological process	Number of genes	Fold enrichment	Raw *P* value	FDR
Regulation of bone remodeling (GO:0046850)	4	23.00	5.23E-05	9.39E-02
Regulation of localization (GO:0032879)	18	2.55	6.12E-05	8.24E-02
Cell migration (GO:0016477)	11	4.22	4.22E-05	9.08E-02
Cell motility (GO:0048870)	12	4.29	1.43E-05	5.12E-02
Localization of cell (GO:0051674)	12	4.29	1.43E-05	7.68E-02
Locomotion (GO:0040011)	12	3.61	7.68E-05	9.19E-02
Generation of neurons (GO:0048699)	12	4.09	2.27E-05	6.11E-02
Neurogenesis (GO:0022008)	12	3.70	5.98E-05	9.19E-02

Table G.5: Complete Gene Ontology (GO) enrichment of community number 13 in the CSD network for bipolar disorder. Note that related classes of an ontology are grouped together. Only results with false discovery rate (FDR) < 0.05 are displayed.

GO biological process	**Number of genes**	**Fold enrichment**	**Raw *P* value**	**FDR**
Myelination (GO:0042552)	4	44.86	2.46E-06	1.32E-02
Axon ensheathment (GO:0008366)	4	44.86	2.46E-06	8.81E-03
Ensheathment of neurons (GO:0007272)	4	44.86	2.46E-06	2.64E-02



NTNU

Norwegian University of
Science and Technology