Co-expression Analysis of RNA-sequence Data from Parkinson's Disease Patients

Akilina Wimalarasan

Department of Informatics-Bioinformatics
University of Bergen

Norway

June 2020

Contents

List of Figures

List of Tables

Abstract

Parkinson's disease is known as a progressive neurological disease characterized by motor symptoms. The motor symptoms are caused by neurodegeneration that causes dysfunctionalities in molecular functions crucial for movement. Network analysis contributes to identifying new biomarkers of diseases by considering the interactions between the disease-specific genes and proteins. This study focuses on a differential weighted gene co-expression network analysis of transcriptomics data, comparing data from healthy persons with Parkinson's disease patients. This analysis method constructs networks and identifies modules that can be compared with different evaluation and analysis methods, to identify dysregulated pathways and causative genes of Parkinson's disease. This disease is a complex disease by multiple variations of symptoms with each individual, hence personalized medicine is highly relevant.

Acknowledgments

First I would like to thank my supervisor Inge Jonassen(*University of Bergen, Department of Informatics, Computational Biology Unit*) for supervising my thesis. I would also like to thank him for regular meetings with close follow up of my progression and interesting discussions.

I would then like to thank Charalampos Tzoulis(*University of Bergen, Department of Clinical Medicine & Haukeland University Hospital, Department of Neurology*) and Gonzalo S. Nido(*University of Bergen, Department of Clinical Medicine & Haukeland University Hospital, Department of Neurology*) for providing me the data and reading through my thesis for quality proofing. A special thanks to Gonzalo for also helping me with basic programming and WGCNA package in R.

Thanks to Christeen Ramanee P. Jesuthasan(*Institute for Cancer Genetics and Informatics*) for reading through the thesis to give me detailed feedback and suggestions that made the thesis even better.

Lastly, I would like to thank everyone else who surrounded me with support and encouragement through the process. I would like to thank my family and my boyfriend for all the support and encouragement despite the distance. And a special thanks to my best friend and all other friends that motivated me through the process. All of them especially helped me through the difficulties caused by the corona situation.

1 Introduction

Parkinson's disease(PD) is a progressive neurodegenerative disorder influenced by both environmental and genetic factors. Age is one of the main risk factors and the average age of onset is 55 years [1]. The symptoms of this disorder are thought to be caused by neurodegeneration of dopaminergic neurons in the brain, which inhibits signaling for movement and makes the daily chores more difficult. Many biological factors are found to be associated with PD: dopaminergic neurons, misfolded proteins, Lewy Bodies(LB), and mitochondrial dysfunctions.

Based on this knowledge a differential co-expression analysis with RNA data from PD patients and controls was carried out to find dysregulated pathways and genes that might underlie PD. The co-expression networks were constructed using weighted gene co-expression network analysis(WGCNA) methodology, where clusters of highly connected genes in the networks are identified and further studied by functional enrichment studies.

The combination of different symptoms with each patient, unknown underlying biology, and the complexity of PD makes personalized treatments(P4 or "personalized" medicine) highly relevant. This study of functionally enriched pathways and causative genes can contribute to the predictive part of P4 medicine.

1.1 Parkinson's Disease

Parkinson's disease(PD) is the second most common age-related progressive neurodegenerative disease after Alzheimer's disease [1], and the most common neurodegenerative movement disorder [2]. The incidents of PD increases markedly with age, where the average age of onset is 55 [1].

Motor symptoms of PD are thought to be caused by neurodegeneration of dopaminergic neurons in the substantia nigra pars compacta (SNpc)(Figure 1). This region of the brain shows more pathological changes with age than any other region [3]. Some characteristics of PD are higher degree loss of dopaminergic neurons and the different patterns of neuronal loss compared to other aging diseases.

Most of the cases of PD are of the "sporadic"(idiopathic) PD, where there is no apparent genetic linkage. Approximately 10% of the PD cases are of the "inherited"(monogenic) form of PD [4]. The patients with "inherited" PD experience symptoms earlier than the patients with "sporadic" PD. The symptoms of early-onset PD(EOPD) are also thought to be caused mutations in the genes which affect protein metabolisms and mitochondrial functions [3]. Studying genetic risk factors in PD has provided insight into possible dysregulated pathways in PD [2].

Pathogenetic study of PD can be studied by transcriptomics(genome-wide expression profiling) [2]. A main goal of transcriptomics study in diseases is to identify differentially expressed genes by comparing multiple samples by using microarrays or RNA-sequencing(RNA-seq). This can then be used to identify over-represented functional pathways that may contribute to the disease process.

At the cellular level, neurodegenerative diseases are characterized by extensive oxidative damage to lipids, proteins, and DNA, which can lead to cell death by a variety of different mechanisms [5]. PD is also characterized by misfolded proteins that lead to toxicity and cell death. There is also a correlation between the age and ability to process the misfolded proteins. Excess misfolded proteins provoke an already compromised proteasome and become a proteotoxic insult to cells [1]. Protetoxicity is described as damage to proteins caused by chemical and physical agents [6].

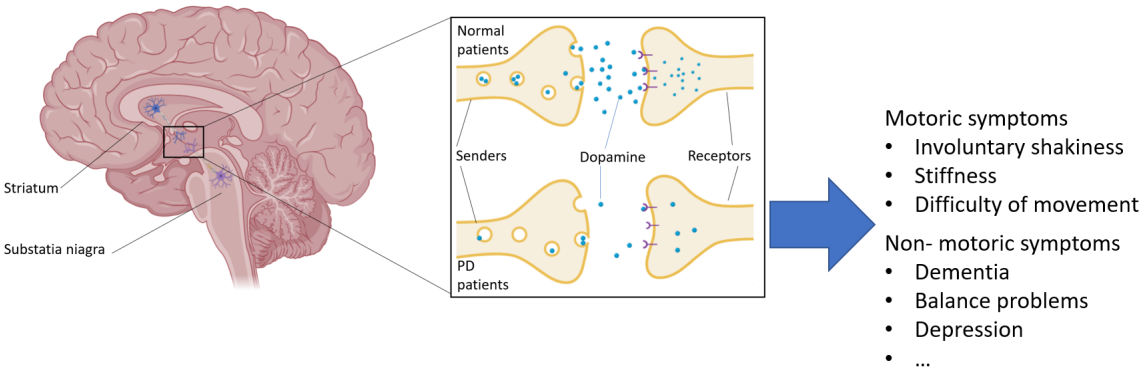

Figure 1: This figure illustrates where neurodegeneration takes place in a brain with the difference of neurotransmitters in the signaling of movement between normal patients and PD patients, and their symptoms are also described. Illustrations created with BioRender.com

1.1.1 Symptoms

Increased levels of dopamine weaken the signals for movement and cause the motor symptoms observed in PD patients. The motor symptoms that characterize PD are involuntary shakiness, stiffness in muscles, and less movement. Tremor associated with PD("Pill rolling") is described as involuntary shakiness at rest that decreases with voluntary movement. This is one of the most common reasons for medical consultation. Rigidity refers to muscle stiffness, for example, an expressionless face. This is caused by increased resistance to the passive movement of the patient's limbs [1]. Other typical symptoms of PD are the difficulty of movement, slow movement(bradykinesia), less movement(hypokinesia), and absence of movement(akinesia). When facing these symptoms daily chores and normal daily routines could be difficult or impossible for PD patients. There are not only motor symptoms, sometimes there can be other dysfunctions giving mental distractions that are not easy to recognize as a symptom of PD. These symptoms are not only caused by the affected regions of the brain but are rather a result of a domino effect from other connected regions that result in balance problems, depression, dementia, sleep disturbance, and loss of smell. Some of the symptoms of PD also occur with advanced aging or other diseases similar to PD, but 80% of the patients with these symptoms manifest for PD [1]. Additionally, it is important to keep in mind that there are no clinical correlations between these symptoms. For example, some PD patients may get dementia before the onset of any motor symptoms. When the motor symptoms appear, however, around 60% of SNpc dopaminergic neurons have already been lost [1].

1.1.2 Genetics of Parkinson's Disease

More than two decades of research have led to the identification of a number of mutations responsible for monogenic and sporadic forms of PD by genome-wide association studies(GWAS). For both forms of PD, the genetic mutations give different age of onset and clinical outcome [7].

Identification of new genes and risk factors linked to PD are found using gene mapping of the human genome and candidate gene approach. Gene mapping is the localization of genes underlying the clinical phenotypes of the disease based on correlation with DNA variants, without the need for prior hypotheses about biological features [4]. New sequencing technology as next generation

Symbol	Disorder	Gene	Discovered
PARK1	EOPD	*SNCA*	1997
PARK2	EOPD	*PRKN*	1998
PARK3	Classical PD	Unknown	1998
PARK4	EOPD	*SNCA*	2003
PARK5	Classical PD	*UCHL1*	1998
PARK 6	EOPD	*PINK1*	2004
PARK 7	EOPD	*DJ-1*	2003
PARK 8	Classical PD	*LRRK2*	2004
PARK9	atypical PD	*ATP13A2*	2006
PARK10	Classical PD	Unknown	2002
PARK11	Late-onset PD	Unknown	2003
PARK12	Classical PD	Unknown	2003
PARK13	Classical PD	*HTRA2*	2005
PARK14	Subtype of EOPD	*PLA2G6*	2009
PARK15	Subtype of EOPD	*FBX07*	2008
PARK16	Classical PD	Unknown	
PARK17	Classical PD	*VPS35*	2011
PARK18	Classical PD	*EIF4G1*	2011
PARK19	Classical PD	*DNAJC6*	2012
PARK20	atypical EOPD	*SYNJ1*	2013
PARK21	EOPD	*DNAJC13*	2014
Unassigned	EOPD	*RAB39B*	
Unassigned	EOPD	*GBA*	2009

Table 1: Classification of genes related to monogenic PD, with a description of what sub-group of PD they are classified into, which gene, and when it was discovered [4] [8] [7].

sequencing(NGS) allows for obtaining whole genome sequences and comparing them to a reference genome. Genome-wide association studies(GWAS) facilitate the discovery of genes associated with human disorders. With GWAS in PD genetic risk factors have been identified in both sporadic and inherited PD by differential analysis of PD patients and healthy persons.

The inherited(monogenic) form of PD is rare inherited DNA variants causing PD with early symptoms. The genes listed in Table 1 are heavily debated as PD is presented to be affected by both genetic and non-genetic factors [8]. These genetic risk factors of the monogenic form are rarely followed when diagnosing PD due to complications of inheritance. In some carriers, the disease will not manifest, and for those where the disease manifests, the age of onset, symptoms, and progression can differ even if it is the same variant within a family [8]. Studies of these monogenic variants could contribute to designing treatments that target a particular genetic cause and pre-symptomatic therapies, and for evaluation of whether these mechanisms are also applicable to the sporadic form [8]. As these monogenic variants of PD are rare, it is challenging to gather such cohorts for studies.

The sporadic form of PD was thought to be caused spontaneously with no association to genetic factors, but as the genetic discovery in PD has increased rapidly over the last decade, the knowledge of genetic risk factors of sporadic PD has grown [7] [8]. PD is now known to be caused by both genetic and non-genetic factors. The genetic risk factors of sporadic PD patients can also be identified by genetic screening and GWAS. Some of the genes identified with monogenic PD are also found in sporadic PD patients, such as *SNCA, LRRK2*, and *GBA* [8]. Sporadic PD has a later-onset of the disease, and when the motor symptoms are observed it might be too late for a treatment. By combining genetic risk scores with observed symptoms, a patient can be diagnosed in an earlier stage of PD.

1.1.3 Transcriptomics in PD

A transcriptome is all the RNA transcripts that express the information content in the DNA of an organism [9]. Transcriptomics has been important in the study of human diseases by enabling the study of gene expression by differential analysis, which may reveal over-represented functional pathways contributing to the disease process [10]. PD is a neurological disease where differential gene expression analysis by transcriptomics of brain tissues may reveal important characteristics of PD can contribute to a better understanding of the disease and improve treatments.

A recent study of transcriptome data in PD has revealed differentially expressed genes and molecular dysfunctions with PD patients, and high similarity with Alzheimer's disease [11]. Borrageiro *et al.* [2] found that the majority of the studies on brain tissues of PD patients used data from the SNpc region and that most of the studies are case-control comparison, but only a minority of the studies used RNA-seq. Among the studies of PD, blood, and skin samples have also been used for studying PD, other than brain tissues [2]. It was found challenging to compare gene lists of studies that examined the same tissue as variations may be caused by sample differences and experimental noise. [2]. Individual studies highlight various genes and pathways, but the most discussed pathways and processes in PD are dopamine metabolism, mitochondrial function, oxidative stress, protein degradation, neuroinflammation, vesicular transport, and synaptic transmission [2]. For further studies, it was recommended that the data from previous studies to be publicly available, consideration of quality parameters, recommendations for statistical parameters, larger sample sizes(minimum 50), and combination of other genomic techniques with RNA-seq [2].

1.1.4 Molecular Factors of Parkinson's Disease

There is uncertainty about the role of toxic environmental and genetic factors underlying sporadic PD. The different factors that are associated with cell loss in SNpc in PD are: misfolded proteins, neuroinflammation, mitochondrial dysfunctions, increased oxidative stress, dysfunctional synaptic transmission, and vesicular transport(Figure 2).

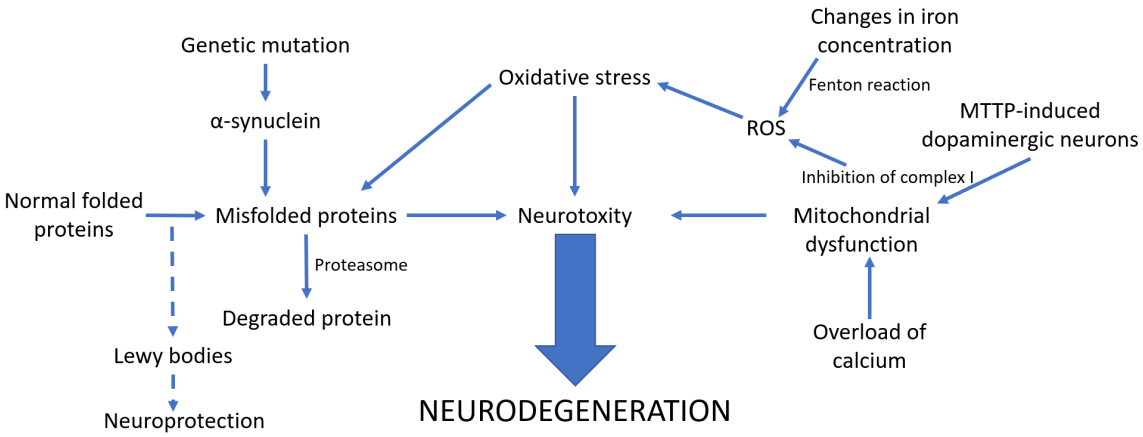

Figure 2: These are some of the multiple factors that are known to cause neurodegeneration in PD patients and how they affect each other. The main factors are misfolded proteins, oxidative stress, and mitochondrial dysfunction.

Studies of the pathogenesis of PD suggests two major hypotheses: (1) Misfolding and aggregation of proteins, (2) mitochondrial dysfunction and consequent oxidative stress [1]. The pathogenic considered by these hypotheses affect different pathways and biological processes. The main goal in current PD research is to understand the molecular interactions and the sequence they act in to gain an improved understanding of PD.

Protein misfolding is a common factor in neurodegenerative diseases. Misfolded proteins can become neurotoxic in many mechanisms, which may cause cell damage by deforming the cell or interfering with intracellular trafficking. Lewy Bodies(LB) are protein inclusions characterized by dark pigmentation and may cause neurodegeneration [12]. LB might also sequester proteins that are important for cell survival, but in Huntington Disease, another degeneration disease, it was suggested by Saudou *et al.* [13] and Cummings *et al.* [14] that there is no correlation between inclusion formation(LB) and cell death. The formation of LB seems more likely to be a defense mechanism against toxic soluble misfolded proteins [1].

Mitochondrial dysfunction is found in both inherited and sporadic PD with evidence of mitochondrial DNA(mtDNA) deletions and decreased complex I activity in SNpc [2]. Studies have suggested an association between correlations of mtDNA mutations and cell loss, also in PD. The double-strand breaks are caused by the damage to mtDNA, likely associated with the highly oxidative environment of the SNpc [3]. This then leads to loss of segments of the mitochondrial genome and further on to reduced mitochondrial function, ATP levels, and proteasomal activity that causes misfolded proteins and eventually cell death [3].

A considerable amount of knowledge in PD originates from the studies of 1-methyl-4-phenyl-1,2,3,6-tetrahydropyrodine (MPTP)-induced dopaminergic neurons. MPTP has shown to block the electron transport chain by inhibiting complex I of the respiratory chain, inducing symptoms of PD [1]. This discovery puts focus on the possibility that an oxidative phosphorylation defect plays a key role in the pathogenesis of PD. Other studies have identified abnormalities in complex I activity in PD and indicated that complex I defect may subject cells to oxidative stress and energy failure [1]. Defects in complex I activity are not exclusively found in the brain but are also found in platels of PD patients [15]. Later studies have shown that decline in the activity of complex IV, also called a respiratory deficiency, may also lead to compromised production of ATP [3]. Respiratory deficiency in complex IV has shown to be caused by a high load of mtDNA deletions within SNpc, higher than other parts of the brain shown in studies of both aging and PD [3]. Dysfunctions of the respiratory chain will affect the SNpc neurons as they are highly energy dependant, and increase their vulnerability.

Reactive oxygen species(ROS) comes from interactions of molecular oxygen with other chemical compounds as calcium and iron, and are more reactive than the molecular oxygen itself [5] [16]. ROS production is unregulated when complex I is inhibited and causes an increase of ROS which leads to cellular damage by reaction with nucleic acids, proteins, and lipids. ROS mainly targets the electron transport chain [17] causing mitochondrial damage and further production of ROS. The presence of ROS increases protein misfolding which in turn leads to the need for the ubiquitin-proteasome system to remove them.

Increased oxidative stress is not only caused by the production of ROS: SNpc neurons are believed to be under additional oxidative stress due to the metabolism of dopamine within the neurons [3]. The metabolism of dopamine(DA) produces hydrogen peroxide radicals and superoxide. By auto-oxidation DA-quinone(a molecule that damages proteins [1]) makes dopaminergic neurons a fertile environment for the production of ROS. Failure of the mitochondrial respiration system may disrupt vesicular storage of dopamine and cause increased DA concentration that leads to cellular damage [1].

DA neurons of the SNpc are also characterized by their pacemaking activity and pigmentation. The pacemaking activity is believed to be important for the maintenance of dopamine levels within the striatum [3] and is maintained by specific calcium channels in adult neurons. The mechanism of maintaining dopamine levels is crucial, and the elimination of toxic factors within mitochondria is key for cellular survival [3]. The calcium channels are also responsible for the modulation of calcium levels within neurons. Overload of calcium may lead to mitochondrial permeability transition, and cause loss of mitochondrial bioenergetic function. Likely for the concentration of iron, where it has been shown an increase of iron in SNpc with age [18] [19] [20] [21] [22]. There are also some contradicting studies of changes in iron concentration with PD patients [23]. Change in concentration of iron causes neuronal loss based on the generation of ROS by the Fenton reaction and will also affect the mitochondria functionality and then cause neuronal loss.

The pigmentation of SNpc neurons is due to the accumulation of neuromelanin [3]. Neuromelanin is a dark pigment composed of proteins, lipids, and products of the DA metabolism, and is thought to protect against oxidative stress [3]. The pigmentation increases with age and has been implicated with cell survival and the loss of neurons in PD, as it is assumed to be a regulator of intracellular iron. A lack of neuromelanin in PD patients than the control group, it is suggested that neuromelanin might protect against neuronal loss caused by intracellular stressors [3].

Changes of synaptic transmission in the striatum are thought to play a role in the occurrence of PD symptoms studied in animal models of PD [24]. It is also thought that synapses maybe

affected at the earliest stages of neurodegeneration [25]. Changes in synapses are found to occur in the striatum as a response to massive dopaminergic loss [24]. Molecular dysfunction of synapses is not only limited to post-synaptic neurons but is also dependent on proper pre-synaptic vesicular transport [24].

Several unknown factors may contribute to cell death in SNpc. Studies in PD patients show that most dopaminergic neurons die long before the symptoms and molecular dysfunctions take place. Motor symptoms, oxidative phosphorylation and ROS abnormalities documented in PD patients could be non-specific features of dying cells [1].

1.2 Co-expression Analysis

Biological systems can be defined by how the molecules interact with each other. Network representations of biological data simplify complex systems and enable the use of various tools from network science and graph theory for data analysis [26]. For example, protein-protein interaction networks are graphical visualizations of which proteins interact with each other to function. Differential network analysis compares topological differences between two different conditions, for example healthy condition vs disease condition.

A key objective of biological research is to identify and understand how all molecules in a living cell interact, and how their functions and interactions relate to the disease [27]. One method to obtain this understanding is by inferring gene function and gene-disease associations from genome-wide gene expression using co-expression network analysis. Co-expression network analysis is a network-based approach that constructs networks of genes based on their correlation in expression [27]. It can be used for candidate disease gene prioritization, functional gene annotation, and identification of regulatory genes. This approach is more effective in the identification of genes that are [27]. Co-active genes are genes that are activated simultaneously, which often indicates that they are active in the same biological processes.

1.2.1 Co-expression Networks

Co-expression networks describe the relationship between genes from their coordinated expression pattern across a group of samples [27]. The construction and analysis of a co-expression network can be described in three steps (Figure 3): (1) gene correlation,(2) network construction, and (3) module definition. The different types of co-expression are signed or unsigned and weighted or un-weighted.

The first step is to identify the pairwise relationships between genes based on their correlation in expressions. The calculation of correlation can be done by methods such as Pearson's or Spearman's correlations which describe the similarity between the genes. In the second step, the construction of the co-expression network is carried out by using the correlation measures. Each node represents a gene and the edges represent the presence and the strength of the co-expression relationship. Finally, the co-expression network is used to cluster the genes into modules, based on connectivity.

Correlation-based co-expression networks use correlation measures ranging from -1 to 1. Unsigned co-expression networks use absolute correlation values, which means two negatively correlated genes are linked by an edge [27]. Maintaining the signed correlation can be problematic when using differential co-expression, since the differences in co-expression between groups that have the opposite signs can cancel out. In signed co-expression networks, this problem of the unsigned network is solved by scaling the correlation values between 0 and 1. Values less than 0.5 indicate the negative

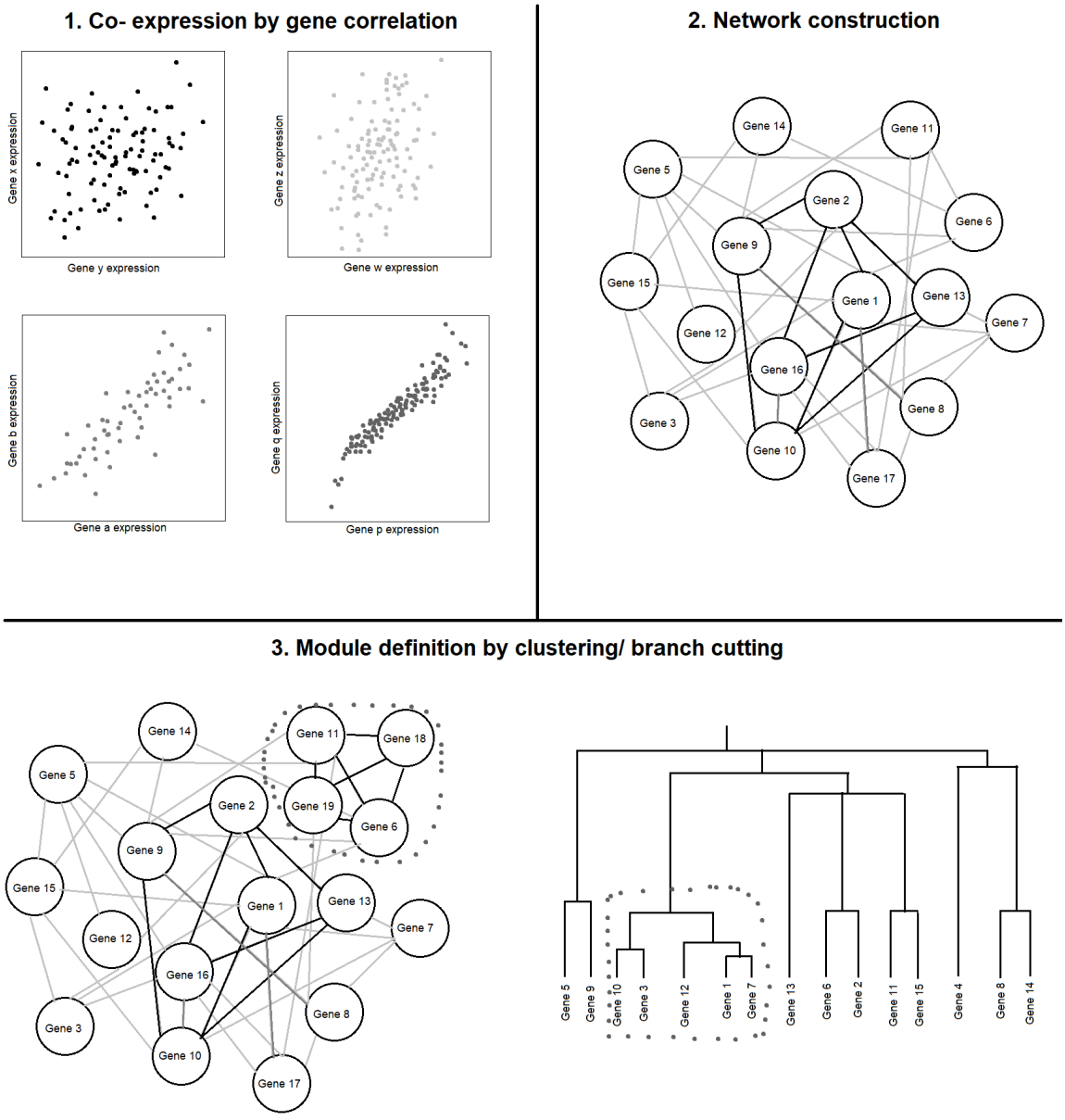

Figure 3: The three steps of co-expression network construction. 1) Calculation of co-expressed genes by correlation. 2) Constructing the network based on the correlations of the co-expressed genes. 3) Defining modules by clustering the network.

correlations and values greater than 0.5 indicate the positive correlations. This scaling constructs a network with more biologically meaningful modules [27].

The edges in the network can also be either weighted or un-weighted. In weighted co-expression networks all genes are connected [27], and each edge is characterized with weights to represent the strength of the correlation between the genes. In an unweighted co-expression network, the weights of the edges are binary(0 or 1), indicating that the genes are either connected or not connected. In unweighted networks, a threshold is chosen so that pairs of genes are defined as "connected" if their pairwise correlation stands above the threshold, or "unconnected" otherwise. A threshold can be set for the edges where hard thresholding gives unweighted networks and soft thresholding gives weighted networks [28]. Till now the focus is more on weighted networks as these construct the most robust networks, and the networks are more informative as it will describe which connections are stronger [29].

1.2.2 Input Data for Co-expression Networks

Data from both microarray and RNA-seq technology can be used for constructing co-expression networks, although RNA-seq data is more used in co-expression studies than microarray [27]. To apply microarray data to create a co-expression network, the probes for all the targeted molecules are required, where this makes limitations of the non-coding genes. RNA-seq data quantifies the expression of non-coding genes in some platforms and has a wider dynamic range that offers a higher resolution for low-abundant transcripts. Many of these non-coding genes play an important role in diseases [27]. Co-expression networks based on RNA data also show an increased resolution for identifying tissue-specific expression patterns and have the potential to differentiate between expression profiles that are closely related [27].

When using RNA-seq data for co-expression networks it is important to obtain expression estimates from the raw sequenced reads, and normalizing and quality controlling the data. The important factors of normalizing are sequencing depth, distribution of counts, transcript length, fragment size, GC contents, and batch effects. There are many different tools for normalizing for each factor mentioned, and new methods are continuously being created to tackle the normalization problems [27]. This process first requires mapping the reads to transcripts, and subsequently count and normalize the counts by the total number of reads(library size). Generally, the resulting transcript quantification is used to test differential expression between groups(e.g. different tissues or different disease conditions). Proper preparation by normalizing and quality controlling the data is crucial for the accuracy of downstream analysis.

To create an RNA-seq-based co-expression network with high performance, a sampling size of 20 or above is suggested based on functional connectivity experiments [30]. Increases in the sample size provide more reliable co-expression estimation, a higher total read depth provides an increase in accuracy, and a higher cut-off threshold may be preferable when data is of higher quality [30] [27].

1.2.3 Network Construction

RNA-seq based co-expression networks are commonly constructed by collapsing overlapping transcript-level expression estimates, and the network is then constructed at the gene-level [27]. This approach does not maintain information about different transcripts encoded by the same gene.

Gene co-expression networks are constructed by describing gene expression profiles in a $n \times m$ matrix where n represents the number of nodes (genes) and m the number of samples. Gene

expression profiles are then used to calculate correlation measures by the pairwise correlations between the genes and stored in a $n \times m$ matrix.

A thresholding procedure is used to create an adjacency matrix based on the correlation values. A *hard thresholding* method gives unweighted edges, defining each entry to be 1 if the similarity measure is above the threshold, 0 otherwise. *Soft thresholding* allows the network edges to take on continuous values between 0 and 1, which results in a weighted network with all the edges connected. Both thresholding methods, however, require the user to set the threshold. The choice of the threshold can be based on how the topology of the resulting network approximates a scale-free distribution [28], or in some cases, a default parameter works as well. Scale-free property distribution is where the distribution of node degrees follows a power law [31]. Once the network is constructed, modules can be identified by clustering the nodes based on how interconnected they are.

1.2.4 Identification of Modules by Clustering

Clustering is used to group genes that have similar expression patterns across multiple samples. The resulting modules often represent sets of genes associated with the same biological processes [27]. Weighted Gene Correlation Network Analysis(WGCNA) is one of the most widely used methods to construct co-expression networks and identify modules using hierarchical clustering. WGCNA clusters genes based on their *topological overlap*, a measure that considers each pair of genes in relation to all other genes in the network, i.e. a high topological overlap between a pair of genes means that they are connected to roughly the same group of genes in the network [27] [32]. Hierarchical clustering based on the topological overlap creates clusters by comparing the genes. The gene profiles are first clustered one by one, then these sub-clusters will be merged and create modules(Figure 4). In the end, all sub-clusters will be connected to one top branch. The length of the branches describes both similarities between sub-clusters, and when the clusters were formed. The similarity is often calculated in the form of distance, as it can be difficult to identify similarity in large data sets. WGCNA has shown to be effective in identifying biologically relevant associations between phenotypes and modules, with both RNA-seq data and single-cell RNA data [27].

The topological overlap measures are used as a similarity metric in co-expression network when identifying modules by clustering. Hierarchical clustering results in a dendrogram, which is then cut at a certain height to define the modules as the resulting branches. To decide the cut height, parameters based on robustness analysis are recommended, but a default parameter could also be used [28].

In the WGCNA implementation, the modules are then labeled by numeric values that are converted to color labels for better visualization. Functional information such as gene ontology(GO) [33] can be used to study the biological meaningfulness of the resulting modules.

are defined as the first principal component of the expression values of the genes within a given module, and they can be considered to be representative of the gene expression profiles within a module [28]. These can be computed by dimension reducing methods and can be used to study the inter-modular relationships by constructing a correlation network of modules, in which the nodes(modules) are linked based on their similarity(correlation between the eigengenes).

1.2.5 Block-wise Network Construction

When it is a large data-set the network construction can be constructed block-wise. To identify the blocks this function pre-clusters the data set with k-mean clustering and merges the smaller

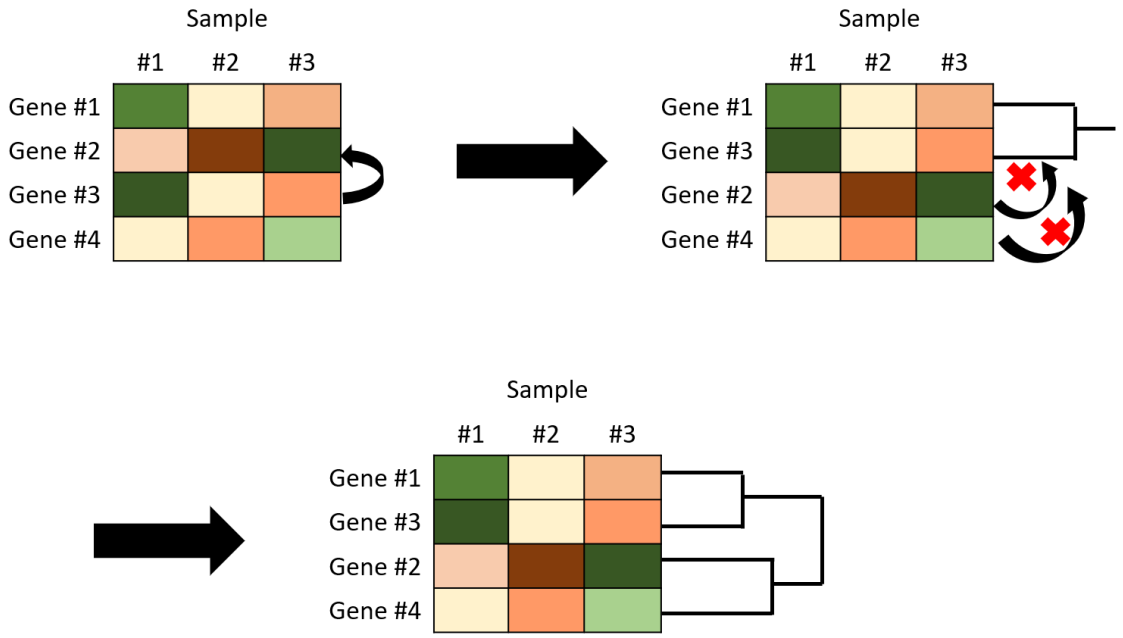

Figure 4: This figure shows briefly how hierarchical clustering creates modules in the dendrogram with a heatmap describing the adjacencies. The columns are the different samples and the rows are the gene profiles. The colors represent the gene expression levels in each sample, ranging from brown to green. By comparing the gene profiles, the gene profiles that correlate the most to each other will make a cluster. In this example, gene profiles 1 and 3 have similar gene expressions in the different samples and are highly correlated. Then the other gene profiles will be compared to the recently created cluster. If the correlation is higher between the pair of genes than to the cluster already made, these two will be clustered together, just like this example. Otherwise, the gene that is most correlated with the cluster will be joined with it to form a cluster of three genes.

clusters to create blocks of defined size [34]. The table describing the block sizes shows a variation of block sizes between the defined minimum and maximum size. Then networks are constructed and modules identified for each block. The time and memory savings of the block-wise approach are substantial: a standard single-block network analysis of n nodes requires $O(n^2)$ memory and $O(n^3)$ calculations, while block-wise approach with block-size n_b requires only $O(n_b^2)$ memory and $O(nn_b^2)$ calculations, analyzing larger data set with blocks of size 7000 feasible on a standard computer [28].

1.2.6 Gene Ontology

There are many databases with integrated data that provide functional information at the gene level. These databases can be used to gain insight into the network modules, by testing their enrichment in specific functional categories [35]. Gene Ontology(GO) is the most widely used annotation database [36]. The GO project integrates data about gene functions from different sources such as research papers, across-species data, and different fields such as evolution and disease studies [36].

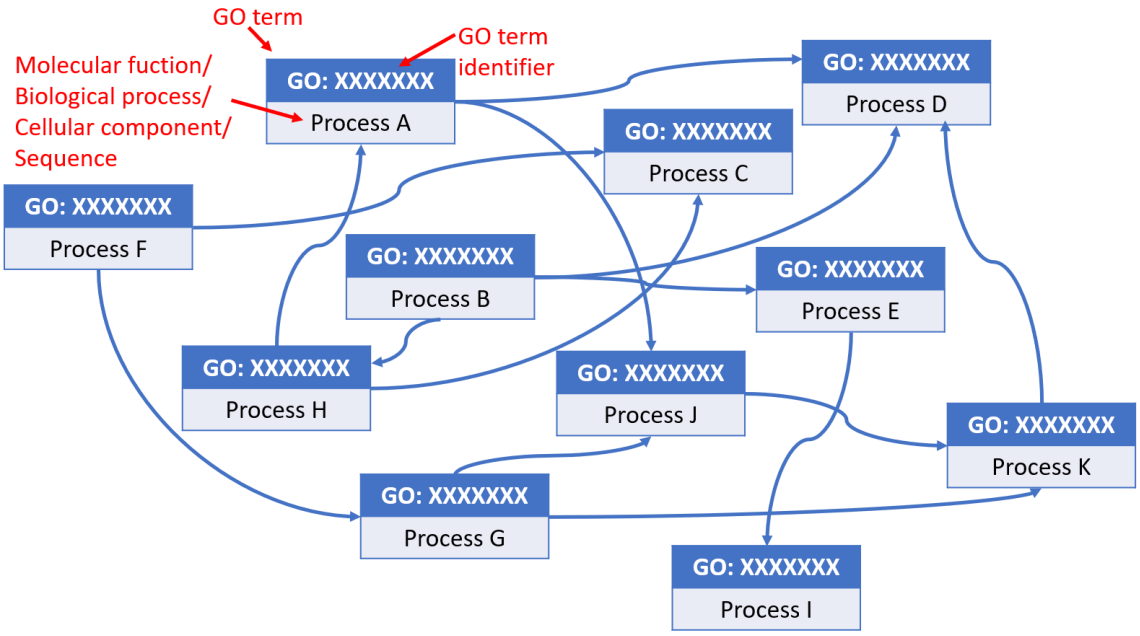

Figure 5: Illustration of GO terms to illustrate how the GO terms define for example a biological process, and how they are connected by their functionality.

GO uses ontologies to describe biological knowledge of gene product functions [37]. Ontologies describe key domains of molecular biology and can be applied in the annotation of sequences, genes or their products. The ontologies are defined in non-overlapping domains and describe the attributes and the linkage between the gene products. GO defines three main domains: "Molecular function", "Biological process", and "Cellular component". The molecular function domain describes activities at the molecular level and the biological process domain describes biological goals accomplished by one or more molecular functions [33]. Cellular component domain relates gene products to their subcellular locations at the levels of subcellular structures including macromolecular complexes. GO has recently added a fourth domain, "Sequence ontology", which provides a classification and a standardized representation for sequences and their features [33]. Gene Ontologies are structured by GO terms(Figure 5) that are related to each other in a hierarchical manner [36]. The mapping between specific gene products and the GO terms are defined as GO annotations. High quality GO annotations are normally based on reviews of published literature and supported by experimental evidence. The GO project is publicly available on their web page and used by many tools to annotate genes with their functions.

1.2.7 Kyoto Encyclopedia for Genes and Genomes

Kyoto Encyclopedia for Genes and Genomes(KEGG) is among the most widely used databases for functional annotation of pathways, along with GO. KEGG is a knowledge-based database for systematic analysis of gene functions in pathways, that links genomic information with higher-order functional information [38]. This database is mainly divided into 3 databases; GENES,

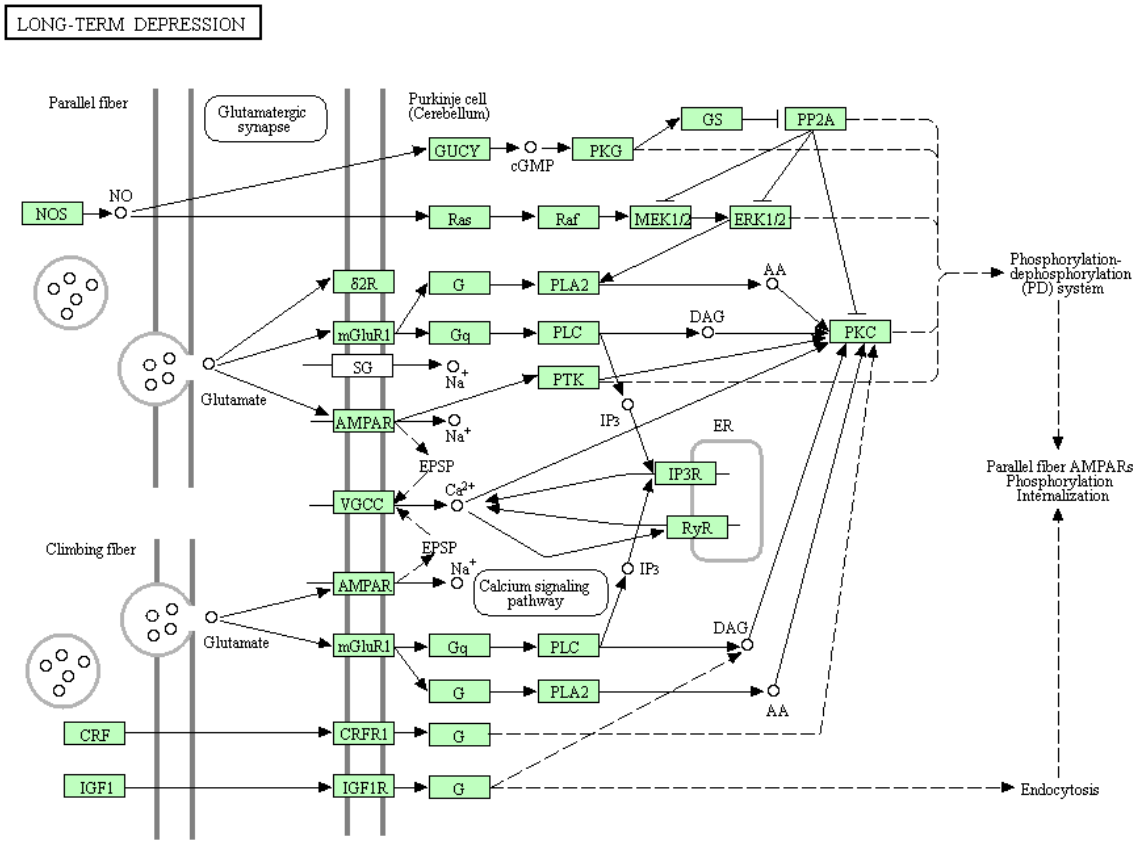

Figure 6: This is an example of a pathway representation of the KEGG pathway "Long-term depression". The network nodes(the green boxes) are the enzymes(gene products), and they are connected by directed edges describing the biological process of the enzymes in this pathway [40]. It also mentions the other pathways they are part of.

PATHWAY, and LIGAND [38], and contains protein-protein interactions, biochemical reactions, gene-regulatory interactions, genetic interactions, and drug-target interactions. Protein-protein interactions represent the largest source of interactions, and genetic interactions and drug-target are the newest added groups, with fewer annotations [39].

The GENES database is a collection of fully sequenced genomes and some partially sequenced genomes with frequent annotation updates. The second database, PATHWAY, contains the higher-order functional information linked to the genes, and in which reactions they are taking part [38]. The PATHWAY database can also integrate graphical representations of cellular processes, where enzymes are represented in a pathway linking their respective identifiers(EC numbers). A pathway is defined as a *reference pathway* when it has been manually validated, and it is used as a template to construct other organism-specific pathways computationally by matching to the EC numbers across species. Figure 6 shows the pathway Long-term depression as a network of enzymes, and some of the pathways it takes part in. Finally, LIGAND contains information about chemical compounds,

enzyme molecules, and their reactions. KEGG provides graphical tools for browsing and comparing genome maps and manipulating expression maps and computational tools for sequence and graph comparison and path computation [38]. For example, with a cluster of genes as a product of network analysis, KEGG can be used to annotate the functions of the genes, and identify pathways that these genes are a part of. KEGG is continuously updated and freely available [38].

The interactions between molecules are important to understand cellular processes. ConsensusPathDB is a database that integrates data from many sources in an attempt to map all the molecular interactions that are known to date. To integrate data from many sources it requires standardized file formats and platforms to exchange data, which can be challenging. During the last years, the number of resources that contribute to ConsensusPathDB has increased rapidly [39], increasing the number of interactions in the database. This database also gives a quality measure of the interactions, as the interactions are collected from different papers and databases of different quality. This database provides a tool that performs a pathway enrichment analysis.

1.2.8 Module Evaluation and Analysis

The purpose of network construction is to identify functional modules that are then evaluated using different methods. These analyses can consist of statistical enrichment of functional terms, comparison to reference networks, or analyses of topological properties.

The functional relevance of a set of genes and the biological meaning of a given module can be assessed by computing the enrichment of GO terms for the group of genes within a module. In each module the enriched GO terms are counted, where the numbers of enriched GO terms vary from module to module [41]. The counts can then be used to avoid some modules with a large number of GO terms, and some modules with few GO terms. It can also be used for over-representation analysis of enriched terms within a gene set.

The functional importance of gene modules can also be assessed by comparing the gene modules in a co-expression network with the structure of a reference network. A reference network can be obtained from biological networks(e.g. PPI-networks, gene interaction networks). Genes within the same module are connected by many high-weight edges, and genes in different modules are connected by many low-weight edges in the reference network.

Preservation statistics can be used to evaluate how well the modules of a reference network is preserved in another network. In a study of gene co-expression modules in type 1 diabetes [42] module preservation statistics were used to evaluate whether a given module defined in the control data set could also be found in the diseased data set. The equations and details for these calculations are described by Langfelder *et al.* [43]. Even though the original paper of $Z_{summary}$ [43] proposes to identify modules with high preservation, Medina *et al.* [42] identified the modules with the lowest preservation with the idea that the weakly preserved modules may highlight the dysregulated pathways in the disease network.

The application of graph theory to the analysis of biological data sets has provided insights into the topology of biological networks [44]. Clustering in a co-expression network often results in large modules, which makes it crucial to identify hub genes, i.e. central genes highly connected with many other genes in the network [27]. Hubs are identified by using statistical indicators of centrality that describes the importance of the nodes based on network topology. Examples of centrality measures are degree centrality, closeness centrality, or betweenness centrality [45].

Topological properties may reveal causative genes when comparing a co-expression network to a reference network. For example, comparing a network constructed based on data from PD

patients to a control network. The study of diabetes type 1 also did a topological analysis with betweenness centrality(BC) measure. Betweenness centrality values indicate the relevance of a node in how capable it is of transferring communication between genes in a module. High betweenness centrality indicates more biologically informative nodes in a module [42]. The genes with the highest BC measures were compared between control and diseased module network.

The topological overlap is another metric that can be used to compare genes between two different conditions. The high topological overlap between two genes indicates that the gene is not likely to be directly involved functionally in the condition, while genes with low topological overlap have activities that are condition-specific [46].

1.2.9 Differential Co-expression Analysis

Differential co-expression analysis consists of measuring the differences in co-expression between different groups(e.g. different tissues, species, cell types, or conditions like healthy vs diseased). Differential co-expression analysis is used to identify biological important differences between modules, and can also be used to identify differences by their topological dissimilarities. Genes that are differentially co-expressed between different sample groups are more likely to be regulators and hence a role in the difference between the sample groups. Some differential co-expression methods do not require the groups compared to be pre-defined [27].

Weighted gene co-expression network analysis(WGCNA) is one of the most frequently used programs for differential clustering based on correlations. It determines the importance of each module in each sample group and calculates an eigengene. The similarity between the eigengenes of the modules can be visualized with heatmaps, that presents the similarities between modules by their eigengenes within a network. By doing this for two condition networks, modules that have strong differences in similarities can be identified as modules containing disease-associated genes. This approach prioritizes which genes are more likely to underlie the phenotype associated with the module.

Another approach to compare networks from two different conditions is to look at which genes are overlapping between modules identified in each network. The overlap between modules can be transformed into a matrix(correspondence matrix), which represents the pairwise similarities between modules using the p-value of a Fisher's exact test.

1.3 Personalized Medicine

Current medicine, how the most of the medicines are designed today, is one treatment for all. Some patients benefit from the treatment, but some patients may not respond to the treatment and show either no effect. Future medicine is treatments tailored to each individual and provides greater chances for a positive effect for all patients. P4 medicine which is another name of personalized medicine consists of 4 parts that need to be considered when designing personalized treatment. These are predictive, preventive, personalized, and participatory medicine(7). P4 medicine is the ultimate goal of systems medicine [26].

According to predictive medicine, with the knowledge of genetic risks associated with many diseases, the signs of symptoms can be recognized before the disease manifests. Unfortunately, the symptoms may happen too late, increasing that the treatment will give no effect or negative effect. Preventive medicine is recognizing earlier signs of symptoms to improve the capacity to prevent the disease. The participatory part is including the patient and the close ones in information about the disease and planning the treatment. Lastly, personalized medicine is the focus of each individual,

by predicting a disease and designing personalized treatment to prevent it. To reach personalized medicine available data and analyzing the data is needed.

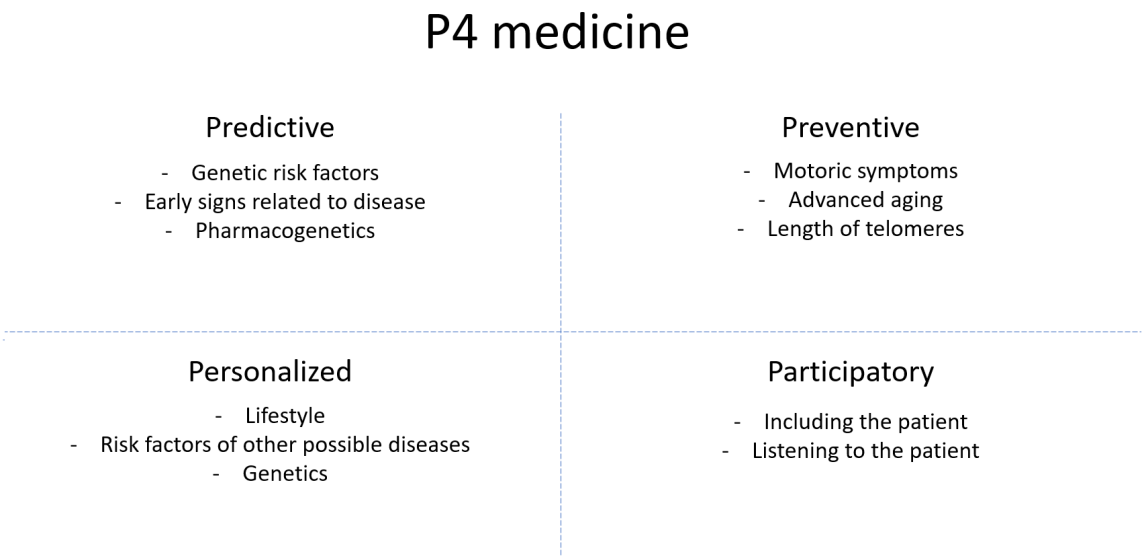

Figure 7: Illustrates P4 medicine. All these factors are equally important to consider when putting together a personalized treatment for a patient. These 4 parts may overlap and depend on each other, like predictive and preventive may be the recognition of symptoms with genetic risk factors.

1.3.1 Personalized Medicine in Parkinson's Disease

The current treatment for PD is levodopa and other dopamine replacement treatment(DRT). These treatments can improve the motor symptoms of PD [47] but do not cure the disease. As PD is a complex disease, DRT as "one treatment for all" is not sufficient. Personalized medicine is an important consideration in PD as each PD patient is different. In PD patients specific personal needs, clinical phenotype, lifestyle, and genetics needs to be considered and may require "cocktail therapies".

In the predictive part of personalized medicine, genes recognized in studies of PD patients could be used for predicting the risk factor of developing PD. Identifying genetic risk factors in the early stage of PD, especially early-onset PD(EOPD), could help precision medicine and prevent the development of PD. The knowledge of genetic risk factors and the mechanisms resulting from the mutation in these genes can be used to develop specific therapies.

Another approach in the predictive part of personalized medicine is pharmacogenetics. Pharmacogenetics refers to the influence of inherited genetic differences in drug metabolic pathways which affect individual clinical responses to drugs as well as adverse events [47]. This approach is slowly evolving and is used in for example studies of the effect of levodopa treatment [47].

Preventive medicine considers the symptoms of a disease at an early stage, often the motor symptoms in the case of PD. As mentioned earlier, the appearance of motor symptoms may be too late, and the neurodegeneration may have caused a significant neuronal loss already. Not only is

PD a complex disease, but aging is also a complex process. Many anti-PD treatments define age as a definitive landmark that influences therapy [47]. When looking for a treatment for a patient, it is important to consider the possibility of side effects, especially if the patient is older. There may be differences between chronological and biological aging in the process of aging, which also makes the treatment decision more difficult. Another factor that may influence personalized and precision medicine is the length of telomeres. Telomeres are crucial for adjusting cellular response to stress as well as the stimulation of cell growth [47]. Accumulation of short telomeres triggers cell death, which makes aging associated with a decline in telomere length.

Personalized and participatory parts of P4 medicine refers to the importance of listening to and including the patient from the very beginning of diagnosis. Each individual has different lifestyles, which should influence affect how the treatment is designed. Not only lifestyle, but each individual is also different in many other ways. Each patient has different genetics and may have other conditions like cardiovascular diseases or risk factors of inherited influence such as diabetes. Cultural background may also be an important factor to consider, so is that the patient is honest and open about themselves and are willing to participate wholeheartedly.

2 Aim of Study

The main aim of this study is to do a differential weighted gene co-expression network analysis(WGCNA) comparing the brain tissue of healthy individuals(controls) and PD patients in order to identify differentially expressed genes.

The sub aims of this study were:

- To apply WGCNA methodology using R packages to the RNA- seq data collected from the brain tissue of healthy controls and PD patients.
- To carry out a differential analysis of the co-expression networks of controls and PD patients and identify and analyze the resulting modules.
- Identify dysregulated pathways within the modules with current knowledge of dysregulations in the context of Parkinson's Disease
- Analyze network topology measures of the modules as sub-networks of the co-expression networks for both control and PD networks, to find genes that might be associated with dysfunctions in PD.
- Identify gene functions associated with dysregulated pathways, critically examine their functions, and relate them to known symptoms or mechanisms related to PD.

3 Methods

The construction of the co-expression network starts with selecting and filtering the data. To identify molecular functions these genes represent, an over-representation analysis of functionally enriched GO terms is performed. By following the WGCNA methodology the co-expression network will be constructed. Module detection is the second part of network construction where dendrograms are created by clustering based on dissimilarity topological overlap matrix(TOM). To merge modules with similar expression profiles module eigengenes are calculated. Heatmap plot is one of the preferred methods to visualize the network. To analyze the results by the modules in the network, the modules are first evaluated and analyzed collectively. Then the interesting modules are identified by preservation statistics and further analyzed to identify causative genes. Most of the methods are based on functions provided in the WGCNA package [28], and the R code is found in the Appendix.

3.1 Data

The data was provided as a count matrix based on RNA-seq expression data from different bulk tissues of the prefrontal cortex. The counts in the matrix represent the count of reads in each sample that overlaps a gene, gene counts per sample. There was a total of 57451 gene-profiles and 123 samples from 3 cohorts: PA-polyA capture RNA seq(74), PW-the Norwegian Park West Study(28), and NBB-Netherlands Brain Bank(21).

The data is first divided into subsets based on the cohorts. For each cohort counts are extracted from the count matrix based on sample identifiers. It is then changed from the data frame to a matrix and transposed so that the rows contain the genes, and each column refers to a sample. Then the data is filtered and quality controlled by removing any genes and samples with too many missing values or very low expression levels. The WGCNA tutorial [28] also recommends clustering the samples to find any obvious sample outliers and remove them. To select the top varying genes, the median absolute deviation(MAD) will be calculated. The data will then be separated based on the conditions and the co-expression networks are constructed for each condition to look at differences between healthy patients and PD patients. An over-representation analysis is done to find enriched GO terms within the most varying genes used in this study.

3.1.1 Quality Controlling and Filtering

Before constructing the network it is important to filter out samples and genes to avoid noise that disrupts the networks. "Good genes" are the genes that pass the criteria of not having missing entries or too low entries. "Good samples" are the samples with the majority of the genes being classified as "good genes". The filtering of "good genes" is done for all the cohorts, and the sample clustering is only done for the selected cohort. The quality controlling and filtering reduces noise and makes it easier to present and analyze the results.

First, the samples and genes with null-values and genes that are not "good" are filtered out. WGCNA package has a function called *goodSampleGenes* that returns FALSE for the gene profiles for each sample that is not qualified as a "good gene". This function checks for missing entries, entries with weights below a threshold that lies in [0,1], and zero-variance genes and returns a list of samples and genes that pass the criteria and are qualified as "good genes" and "good samples" [28] [34].

The second step is clustering based on samples to identify any obvious sample outliers to remove them. Here the samples are clustered by hierarchical clustering and then the sample tree is plotted

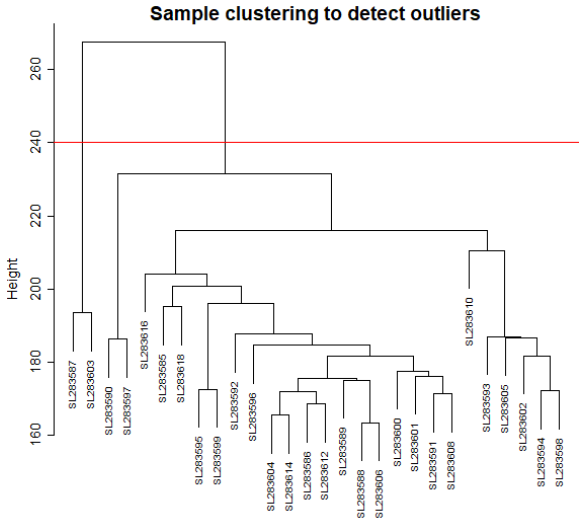

Figure 8: Illustrates a sample clustering tree where the two samples to the left are obvious outliers, and therefore the tree is cut at the height 240 and excludes those sample outliers.

to identify the outliers. A sample outlier is identified by looking at the sample tree to find the sample(s) that is distant from the other samples in the sample tree. Figure 8 shows that the two samples to the left are outliers from the rest of the samples that lie under one branch of the sample tree. To remove the sample outliers a cut off height is set and the cluster of outliers is separated from the rest that should be kept. In Figure 8 the cut height is set to 240 cutting away the 2 samples at the most left and the rest are kept. The filtering of genes and sample outliers reduces possible noise for creating networks and clusters.

The most varying genes can be filtered out of a large set of genes, which will also be the most differentially expressed genes between the conditions. After removing the genes and samples that were not qualified as "good", there might still be a large set of genes in the count matrix. The most varying genes can be filtered out by calculating the median absolute deviation(MAD). These values are calculated by the function *mad*. Median absolute deviation takes each gene as a vector and calculates the variation by the formula

$$constant * cMedian(abs(x - center))$$

where *constant* is default value = 1.4826, and *center* being the median of vector x [48]. Higher MAD means higher variation among the samples. The most varying genes are selected by sorting the list of MAD values in descending order, matching the gene identifiers in the list of MAD values with the gene identifiers in the original count matrix, and then the top varying genes are used in the final count matrix.

This subset with the top varying genes by the calculated MAD values is then divided into PD patients and healthy persons. For the selected cohort, one network will be constructed per condition. For the analysis, the reference network will be the network created from control data. The separate network construction will then be compared through different analysis methods to identify differences in pathways and gene functions.

3.1.2 Over-representation analysis of Gene Ontologies

The over-representation analysis is used to identify over-represented GO terms within a gene set and will describe the molecular processes that the genes present. It will also indicate which processes are more significant in the gene set and can be used to give an idea of what molecular processes the most varying genes between PD patients and healthy persons are associated with. The GO terms and their p-values identified with GO::TermFinder are then visualized with REViGO, a tool for summarizing and visualizing long lists of GO terms with calculated p-values.

GO::TermFinder [49] is open-source software that takes a gene list as an input, and the ontology domain and a reference set with annotation data are selected. The numbers of annotations to a GO term are compared between the input list and the reference set. This software calculates a p-value using the hypergeometric distribution that represents the statistical significance of a GO term associated with a group of genes in the input list [49]. This calculation considers the total number of genes estimated for an organism in the annotation data, and the number of genes within that organism having that GO term annotation [49]. The higher these numbers of genes are, the closer the p-value is to zero, which indicates more significance of a GO term in the user-list of genes [49]. This software also calculates correction for multiple hypotheses by Bonferroni correction and false discovery rate, and the last measure calculated is false positives [49]. A gene product can be annotated to one or more GO term(s) and is then also annotated to the related GO terms of the connected GO terms [49].

The output is a table listing of all annotated GO terms within the selected domain with the p-values, results of multiple hypotheses testing, and false positives. The resulting list also describes the numbers of genes from the input list that are associated with a GO term and the number of genes from the annotation list that is associated with this GO term.

The results of the identified GO terms can be visualized in REViGO(Reduce +visualize Gene Ontology) [50]. REViGO is a web server that uses clustering to summarize and visualize long lists of GO terms in multiple ways such as table format with hierarchical description, scatterplot, graph-based visualization, treemaps, and tag clouds.

This software takes the list of GO terms with their p-values from GO::TermFinder with settings of semantic similarity, a description of the numbers provided in the list [50]. Semantic similarity describes how similar the GO terms are based on their associations with other terms and how similar the processes or functions are [50]. A lower cutoff value for semantic similarity will give a shorter output list, which can lead to removing GO terms without statistical support [50]. The algorithm removes functionally redundant GO terms by calculating semantic similarity, clusters highly similar GO terms and finds a representative GO term for each cluster guided by the associated p-values from GO::TermFinder [50].

The scatterplots visualize the semantic similarities of the GO terms by placing the more similar GO terms closer to each other in clusters, by composing eigenvalues of the terms' pairwise distance matrix [50]. Then a stress minimization step improves the agreement between the GO terms' semantic similarity and their closeness in the plot [50].

The graph-based visualization presents the GO terms as nodes and only 3% of the strongest GO term pairwise similarities are designated as edges in the graph [50]. Only the strongest edges are visualized and the threshold is found by balancing over-connected graphs with no visible subgroups and very fragmented graphs with too many small groups [50]. The layout algorithm used in this software is ForceDirected layout [50], which describes the similarity of the nodes by distance [51]. In both scatterplots and graph-based visualization, the size of the circles indicates the generality of the GO terms, and the p-values are described by their colors [50].

The last two visualizations are treemap view and tag clouds. The treemap view illustrates the hierarchy of the GO terms with the cluster representatives [50]. The tag clouds show over-represented keywords in the GO term's descriptions of the GO terms in the input list [50].

3.2 Network Construction and Module Detection

Network construction is the second step of WGCNA methodology [28] of co-expression analysis. To construct the network a co-expression matrix is necessary, which will be modified into an adjacency matrix by setting a threshold value to highlight the stronger similarities. To minimize the effects of noise, the adjacency matrix is transformed into a topological overlap matrix(TOM) and the corresponding dissimilarity matrix is calculated. When the network is constructed the next step is module detection by clustering using the TOM.

3.2.1 Threshold

A threshold can be set to highlight the strongest connectivities for further analysis of the co-expression networks. The thresholding procedure modifies the adjacency matrix, either by setting a threshold and removing the weaker linkages below a value(hard thresholding) or by raising all entries to power highlighting the strong connections(soft thresholding).

Weighted gene co-expression network provides the most biological meaningful co-expression network. To get a weighted co-expression network a soft-threshold is set. The first step is to choose a set of soft-threshold powers, and then plot them by calling the network topology analysis function *pickSoftThreshold* which will do the scale-free topology analysis of the set of powers [29]. To identify a threshold value that satisfies the scale-free property, the scale-free topology fit index is plotted with a red line indicating the powers that satisfy the scale-free property. The powers close to and above this line satisfy the scale-free property. The power that will be chosen is the lowest integer that can satisfy the scale-free property, as proposed by Medina *et al.* [42]. From the network topology analysis in Figure 9 the power 3 is is the lowest power above to the plotted line, and therefore the soft-threshold value can be set to 3.

3.2.2 Topological Overlap Matrix

The adjacency matrix A is a symmetric matrix with entries a_{ij} describing the co-expression similarity by correlation measurements, which encodes the connection strength between nodes i and j. The correlation measures in the WGCNA package are calculated by Pearson's correlation measure method that lies in [-1,1]. These correlation measures are then scaled to lie in [0,1], which results in a signed network. This adjacency matrix is then modified by the set threshold power, which transforms the adjacency matrix into a topological overlap matrix(TOM).

Topological overlap measures describe pairwise interconnections and can be used to identify modules. The topological overlap between the nodes i and j reflects their relative interconnectedness, by measures that lie in [0,1] [52]. The measure equals to 1 by two conditions; one is that all of i's neighbors are also j's neighbors and two is that i is connected to j. Topological overlap equals 0 if both nodes are unconnected and they do not share any neighbors. The TOM can be considered as a "smoothed out" version of the adjacency matrix [52].

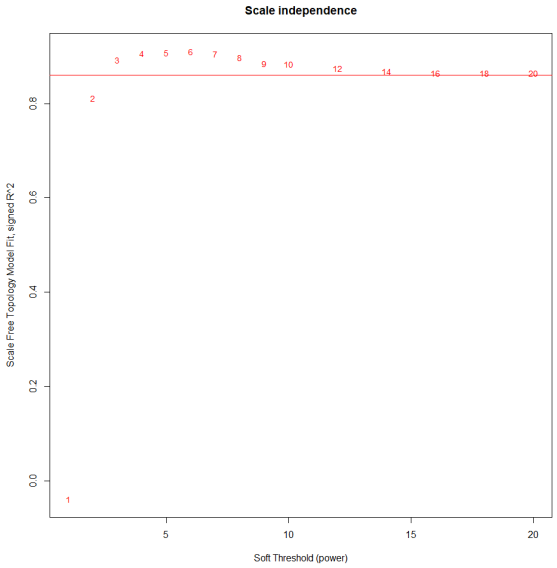

Figure 9: Example of a plot analyzing network topology for various soft-thresholding powers. This plot describes the scale-free fit index (y-axis) as a function of the soft-thresholding power (x-axis) [28].

3.2.3 Modules

Modules are defined as clusters of densely interconnected genes [28]. The topological overlap matrix(TOM) is used to identify the modules by hierarchical clustering done by the function *hclust*. This function performs a hierarchical cluster analysis using the dissimilarity matrix for the n objects being clustered. At each stage, cluster distances are recomputed according to the clustering method being used [28] [34]. This clustering creates a dendrogram(Figure 11), that describes the clusters in a tree. Each leaf corresponds to a gene, and the branches of the dendrogram group together densely interconnected highly co-expressed genes.

Branch cutting methods can be applied to the hierarchical dendrogram to identify modules, as the branches of the dendrogram correspond to modules. The modules are identified by the dynamic tree cut method that performs a branch cutting to detect modules of a minimum size specified and a variable of *deepsplit*. The deepsplit variable provides control over the sensitivity of cluster splitting, where a higher deepsplit value gives more modules of smaller size [53]. The function outputs a vector of numerical labels giving an assignment of objects to modules [53]. Module 0 contains the unassigned genes, and the rest of the numeric labels describes the size of the module where the module 1 is the largest, and module 2 is the second-largest module and so it continues. The module labels are then converted from numeric values to color labels, which can be plotted together with the dendrogram, which is the upper color-range in Figure 11. The grey module indicates the genes in module 0, unassigned genes.

Modules eigengenes are used for summarizing the gene profiles of an identified module. The module eigengenes can be calculated by the function *moduleEigengenes* [28], which takes a count

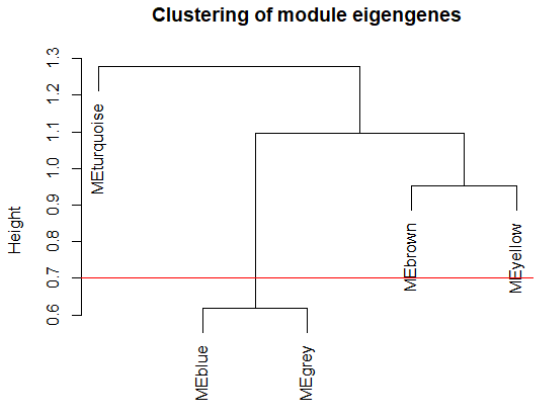

Figure 10: This is an example of a module eigengene tree with a red line describing the modules to merge. The module eigengenes below the red line are the ones that will be merged.

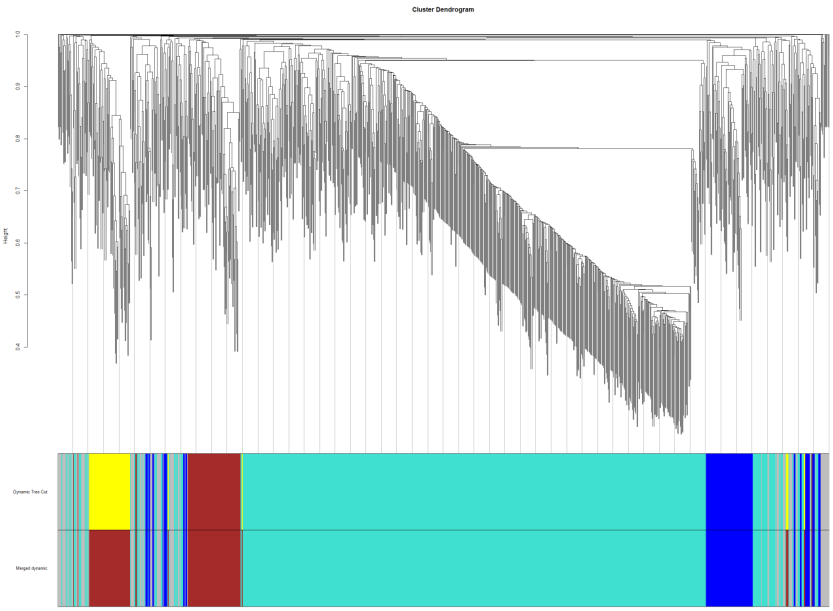

Figure 11: This is how a dendrogram will look for 1000 genes, where the upper colors explain the modules before the merging and the colors below describe the modules after the merging.

matrix and the colors from the modules in the tree as parameters. Then the dissimilarity of module eigengenes are calculated and the module eigengenes are clustered by hierarchical clustering using the dissimilarities. With the tree of module eigengenes, a cut is performed by defining a cut height that describes which modules to merge(Figure 10). The modules with similar expression profiles are merged by an automatic merge function *mergeCloseModules*, that uses the correlation of the eigengenes to measure similarity used for merging the close modules [28]. The merged modules are converted from numeric to color labels again, and the dendrogram will be plotted with both color-labels of the modules, before and after merging the modules to see what the merging did to the modules. Figure 11 shows the merging of the yellow and brown module, which indicates that they were modules with similar expression profiles.

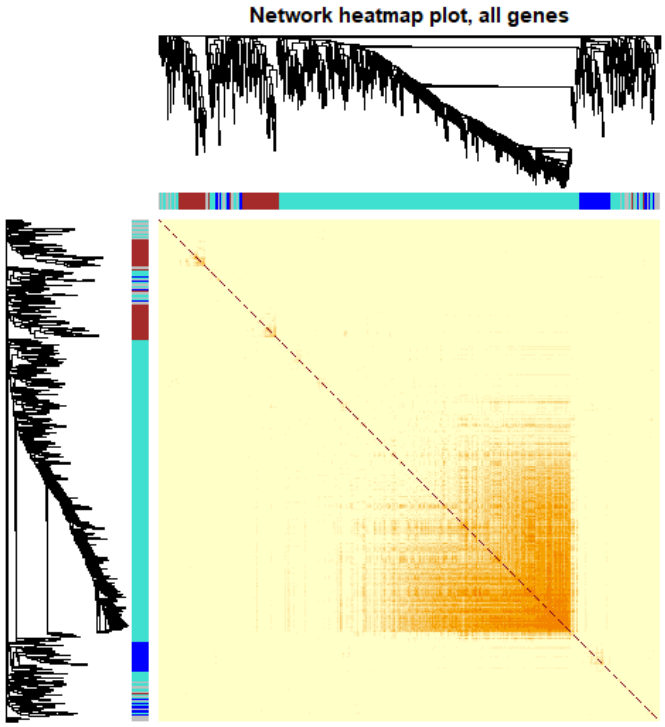

Figure 12: An example of a network visualization of the dendrogram in Figure 11 by heatmap. The darker the color is, the higher is the adjacency.

3.3 Visualizing

The dendrogram created by the topological overlap matrix and the modules labeled by colors will not give a clear overview of the network for analysis purposes as seen in Figure 11. The TOM can be visualized by a heatmap plot(Figure 12) o visualize the dissimilarities. This will describe the connections by a color-code indicating strong and weak connections between the gene profiles. The connections are calculated by transforming the dissimilarities of the TOM with a power, which

makes the moderately strong connections more visible in the heatmap. Each row and column of the heatmap correspond to a single gene expression profile. The lighter shades indicate low adjacency and the darker colors denote higher adjacency, where adjacency describes the similarity between the gene expression profiles. The gene dendrogram and the module colors will also be plotted along the top and left side of the heatmap, as in the example shown in Figure 12. Blocks of darker colors along the diagonal represent the modules. For a more informative plot, the diagonal of the matrix is set to not applicable(NA). The heatmap is then plotted by the function *TOMplot* which visualizes the heatmap plot with the TOM, dendrogram, and module colors as input values.

3.4 Evaluation

An evaluation of the modules identified for both diseased and control networks will reveal interesting modules that can be analyzed further to identify the dysregulated pathways that might contain causative genes associated with PD. In this study module eigengene heatmaps, correspondence matrix and module preservation statistics are used to evaluate the modules, and for determining which modules to analyze further. Module eigengene heatmaps are used to evaluate the modules by describing the adjacencies of the module eigengenes within a network. Correspondence matrix compares two networks to visualize the count of genes per module, the overlapping genes, and the similarity between the modules of both networks. Module preservation statistics visualization describes how well the modules of a reference network are preserved in a test network.

3.4.1 Module Adjacency Heatmap

The eigengenes can be used to describe module similarity by eigengene correlation to evaluate the modules by studying the relationships among the found modules within a network [28]. The dendrogram of module eigengenes can be used to plot a heatmap describing the relationships between them by using the function *plotEigengeneNetworks*. These relationships are visualized by a color-code from blue to red, where red indicates high similarity, and the blue indicates low similarity. In this method, the similarity is based on the correlation of the modules by their module eigengenes.

As seen in Figure 13, the module eigengenes indicated by color labels are plotted with a heatmap-plot visualizing the similarities of the module eigengenes. High similarity indicates similar biological processes, as a module presents the genes with similar associated biological processes. The blocks in the heatmap define groups of correlated module eigengenes and are called meta-modules, which represents similarity in biological processes between these modules.

3.4.2 Correspondence Matrix

The module correspondence matrix is a method to relate modules from different networks to each other. A correspondence matrix is created by comparing two sets of modules and counting overlapping genes between the corresponding modules. The indices with the overlapping counts between the modules are colored from white to red, where red indicates a higher similarity between the modules of the data sets. The analysis is presented by the module eigengenes, module labels and colors, and the tree.

The tables of p-values and the counts are filled by pairwise comparison between the module eigengenes for both data sets. The p-values are computed by Fisher's exact test, also known as the hypergeometric test, and describes the overlap of gene expression profiles of two modules. The smaller p-values are cut off to add the color-code and display the tables in a more informative

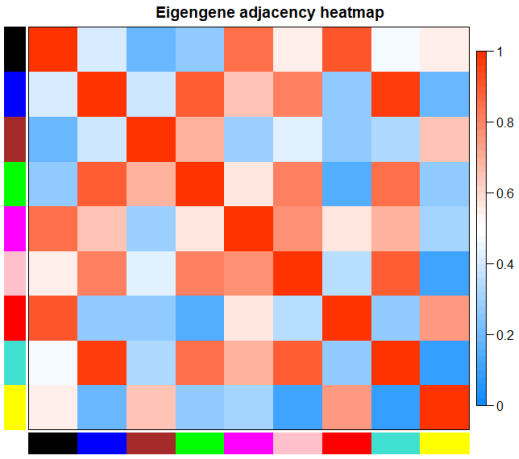

Figure 13: This is an example of a module eigengene heatmap describing adjacencies between the modules within a network. Red indicates more similarity and blue indicates less similarity. For example, the dark blue module is highly similar to the turquoise module, and the black and red modules as well. The turquoise and yellow modules show very low similarity.

way. The matrix is displayed by the function *labledHeatMap* where the coloring from white to red encodes $-log(p)$, where p is the calculated p-value [28]. The rows and the columns represent the modules of each data set, with the network name included in the labels.

3.4.3 Module Preservation Statistics

Module preservation statistics are used to evaluate the modules identified in a reference network against the modules in a test network. Module preservation is calculated by the function *modulePreservation* in the WGCNA package, where the input is a multiset of the reference and test networks with their modules.

Module preservation statistics calculates how well the modules of the reference set are preserved in the test set, in a pairwise manner [28] [43]. This function uses the gene names for matching, so the column names must be valid and at least half of the genes in both sets should match for evaluating the module preservation [28] [43]. The module preservation results in a nested list containing statistics describing quality, preservation, accuracy, reference separability, test separability, and permutation details. For further analysis the observed values and their preservation scores(Z_{scores}) are isolated. The Z_{scores} are summarized to $Z_summary$ values that are visualized to find interesting modules, as in diabetes study [42]. In the article about network analysis of type 1 diabetes [42], it is suggested that the modules with the lowest preservation(lowest $Z_{summary}$) are the modules containing the dysregulated pathways.

3.5 Analysis of Interesting Modules

This study focuses on identifying gene products that could be a genetic factor of PD patients, which could improve the treatments by better knowledge for predicting the disease. The gene

products can be identified within dysregulated pathways present in the modules. The evaluation of the modules reveals differences and similarities between the modules. This can be used to identify interesting modules that can be analyzed further to find causative genes and dysregulated pathways. Interesting modules could be identified by combining the evaluations of the adjacencies in the module eigengene heatmap, overlapping genes or similarities in the correspondence matrix, and the modules that are preserved at a low level in the preservation statistics. Each module is a set of genes, and the genes can be plotted into tools that use pathway databases with GO terms to identify the gene functions and their associations between each other, and what pathways these genes are present in. Further on a comparison of identified genes and pathways can be performed to analyze the differences between the PD patients and healthy persons. The dysregulated pathways can be found by tools providing over-representation analysis of the pathways present in a gene set. Network topology measures may also reveal differences in gene products when comparing between control modules with case modules.

Many of these tools and databases are available online, with a quality measure which indicates how reliable the results are based on what resources the information was fetched from. In this study ConsensusPathDB is used as proposed in the study of diabetes [42], which includes the enrichment of predefined pathways by KEGG and GO terms. Cytoscape is used to visualize the module networks for analyzing network topology measures such as betweenness centrality and node degree. Cytoscape also provides plug-ins where ClueGO is used in this study to visualize the interactions of the pathways within a module.

3.5.1 ConsensusPathDB

ConsensusPathDB is an online tool for analyzing sets of genes by different analyzing methods [54] [55]. In this study, an over-representation study is used to identify pathways with the genes in the interesting modules. The input is a set of genes together with different settings like a database of pathways, p-value threshold, and minimum overlap size. The minimum overlap size tells how many genes from the gene set should match with the gene set of a pathway to include that pathway in the output list.

The output is a list of pathways with their set size, overlap size, p-value, q-value, and pathway source. This list can be downloaded in a tab-separated file, that contains the p-value, q-value, pathway name, source name, external id(source id), gene symbol of overlapping genes, and their numeric ids and overlapping size. With the output list online it was possible to generate a word cloud that shows which words appear the most of the pathway names. The p-value is calculated by the hypergeometric test considering the number of genes present in the user-specified list and the annotation data [54] [55]. The q-values describes the correction of p-values by multiple testing using the false discovery rate [54] [55]. The downloaded list can be used to identify the matching genes for each pathway. The most interesting genes will be the genes that match with pathways that are found to give outbreaks of the disease. These genes might be in the control module as well, but it depends on how many genes are set to the minimum overlap size in the pre-settings.

3.5.2 ClueGO-Network of Pathways

Cytoscape is a software for visualizing molecular interaction networks and biological pathways and integrating the networks with annotation data [51]. Cytoscape provides the basic functionality of integrating data on the graph, visualization, selecting and filtering tools, and implemented external methods as plug-ins.

ClueGO is a plug-in app in Cytoscape that reads a gene set and visualizes the network of pathways by selected data integration. This app provides multiple analysis methods, and in this study, the analysis method is set to functional analysis. A gene list is provided from each module as an input. This tool also allows selecting between different databases, and if more databases are selected they can be separated by the shape of the nodes. In this study, KEGG is used to make the results comparable with ConsensusPathDB. The network specificity is a range of how detailed the network should be in the visualization. This is a toggle bar that can be moved from global to medium to detailed. A threshold of the p-values can also be set as in ConsensusPathDB.

The other settings such as statistical and grouping options are set to default, and the default for preferred layout is "Prefuse Force Directed Layout". The layout can be selected to groups, which calculates similarities and clusters similar pathways into functional groups. The functional groups are shown by colors in the final network. The size of the node represents the node significance, and the node-label is highlighted by the same color as the nodes of a functional group present the representative pathway of a functional group [56]. The representative pathway is selected based on the highest percentage of genes per term in the functional groups. Some pathways are a part of more than one functional group, then the node color will be like a sector diagram and present all the functional groups.

The edges are weighted by a Kappa score that shows how similar their associated genes are and defines the connectivity between them. The Kappa score is calculated by Cohen's kappa coefficient measure, [56]:

$$K = \frac{p_o - p_e}{1 - p_e}$$

Figure 14 describes the formula in more detail and shows an example of how to calculate the kappa score by Cohen's kappa coefficient measure. First, a binary matrix with the pathways and genes is calculated by setting 1 if the gene is in the pathway, 0 otherwise. Then a Kappa score is calculated between all terms(pathway) creating a term-term similarity matrix by counting genes in both pathways and classifies as "yes" or "no". The "yes"-"yes" column describes overlapping genes between the two pathways which is used to calculate p_o, and "yes"-"no" column describes the genes that are in one of the pathways and not in the other, and is used to calculate p_e with "no"-"no". The p_o is the relative observed agreement among the pathways, a ratio of overlapping genes among the genes present in both pathways. The p_e is the hypothetical probability of chance agreement, which measures the probability of the genes which are only found in one pathway to be in the other pathway for both pathways [56]. These two ratios are then used in the Cohen's kappa coefficient measure which calculates K-Kappa score.

Figure 15 illustrates the results produced by ClueGO. The nodes present the pathways and the edges describe the connectivity by the Kappa score defined by shades of grey, where darker shade indicates a higher score. In some networks, often more complex networks, the nodes are a part of several functional groups, and the nodes are then colored like a sector diagram to present all the functional groups of the pathway. This plug-in to Cytoscape has almost the same functionality as ConsensusPathDB, but because of the network visualization of the pathways, it is easier to see the similarity between the pathways.

3.5.3 Module network visualization

These interesting modules can also be visualized in Cytoscape where the genes and their interaction can be analyzed. This tool allows for a variety of graph layout, where spring embedded layout is set

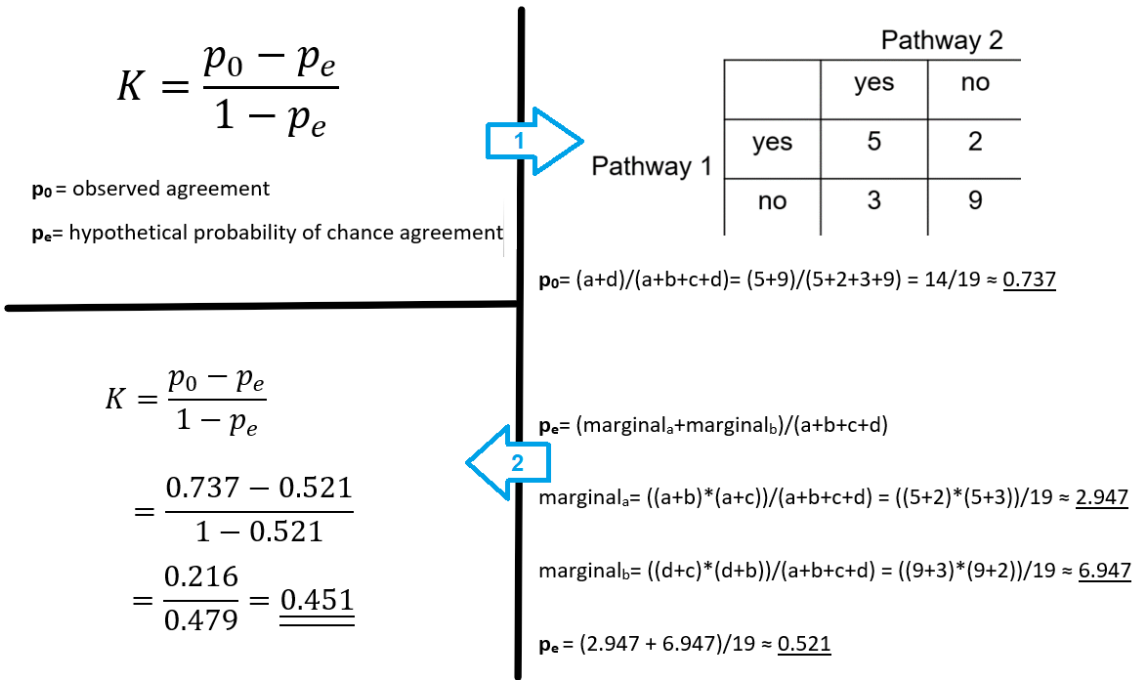

Figure 14: The kappa score is calculated by Cohen's kappa coefficient measure which describes the similarity between two pathways. Here is a simple example of how it is calculated, where the matrix is a count matrix of genes in both pathways. "yes"-"yes" indicates overlap of genes, "yes"-"no" indicates there is a chance for overlap, and "no"-"no" indicates no overlap nor a chance for overlap [56]. In this example, the pathways have 5 genes in common, and pathway 1 has 2 other genes as well and pathway 2 has 3 other genes. Both pathways do not have the 9 other genes, that are found in other pathways of the entire network.

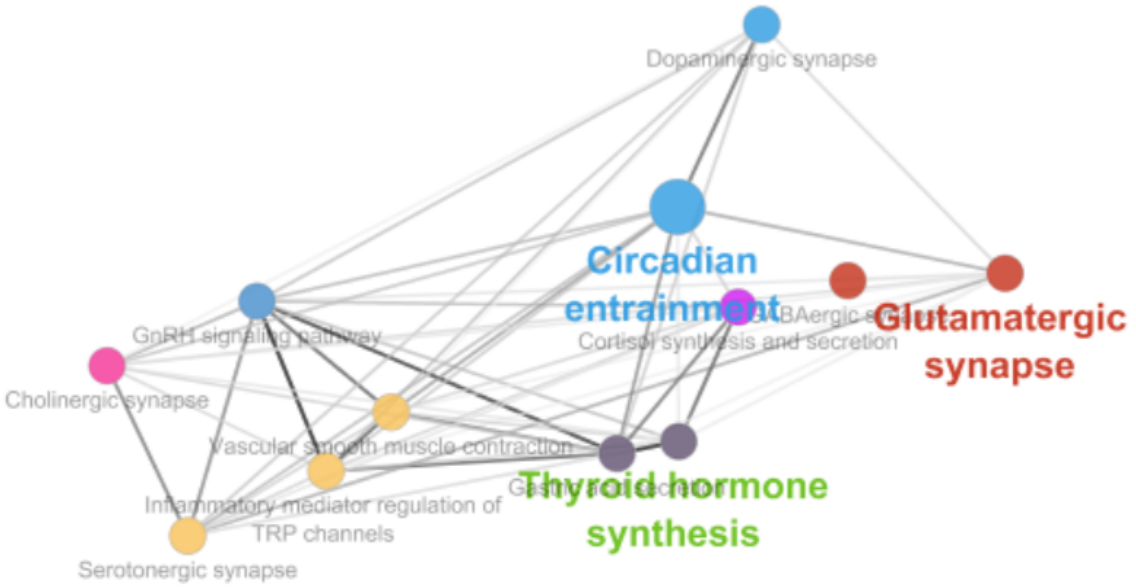

Figure 15: This is an example of a pathway network with KEGG data in ClueGO analysis. The nodes are the terms(KEGG pathways), and the edges are weighted by a Kappa score. The color of the nodes indicates the functional group, which are highlighted in the same color.

to default and is the most widely used layout in Cytoscape [51]. The layout models the nodes by similarity, which means that the nodes closer to each other are more similar than the nodes further away. Other attributes can be applied to the layout for the style of the nodes and the edges.

For the modules that appear to be very large, the analysis can be complicated without filtering the nodes and the edges within a module network. The filtering can be done in Cytoscape based on the attributes, or it can also be done when extracting the module matrix by selecting top-ranked genes by calculating soft-connectivity in the topological overlap matrix(TOM) for each module, and by setting a threshold for the weights. The edges are weighted by similarity measures from TOM, and by setting a threshold the number of edges is limited to only the edges with a similarity measure above the threshold. This gives us only the strongest edges and makes the visualization of the network less complicated. The edge file is used in Cytoscape by defining the source and target column, and the node file can be added if there are some extra attributes of the nodes.

The network analysis tool in Cytoscape gives statistical measures of the network topology such as node degree, betweenness centrality, and clustering coefficient. These networks and topology measures can also make it easier to compare genes in two modules and look at the different measures between control and disease.

The node degree can be used for coloring or sizing of the nodes, which makes it easier to compare overlapping nodes(genes) between control and case module network by their connectivities. As in the diabetes study by Medina *et al.* [42] where the betweenness centrality measure of nodes in control and case network are compared, other measurements can also be compared for each overlapping node. The number of overlapping genes will vary according to the set parameters of the module networks when writing to the files.

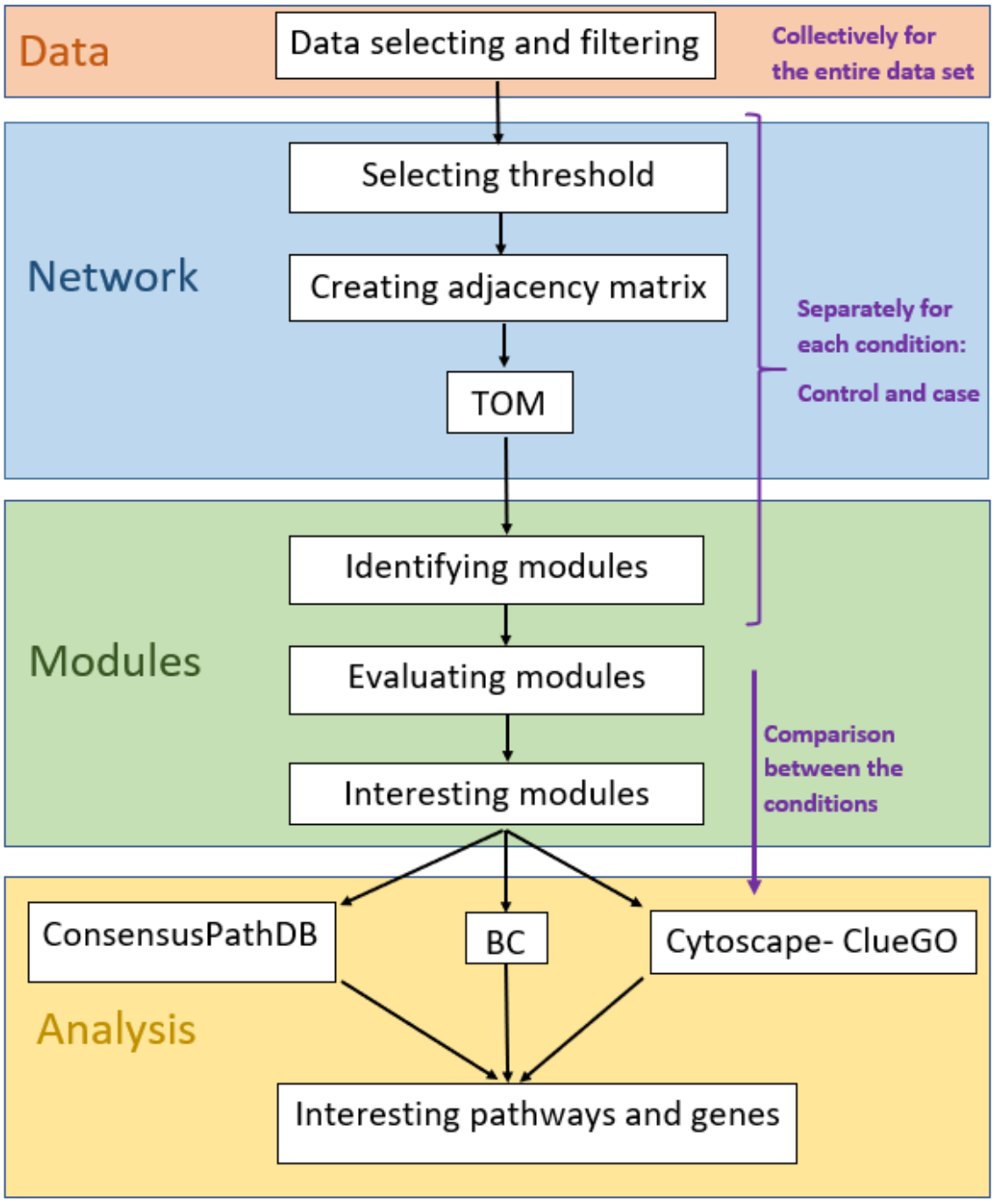

Figure 16: This figure summarizes the flow of the co-expression analysis performed in this study. The purple text indicates how the data is used collectively in filtering, then separated for network construction and module identification, and compared in the end for comparison.

4 Results

The aim of this study is to perform weighted gene co-expression network analysis(WGCNA) with data from PD patients vs healthy persons, to identify dysregulated pathways and causative genes associated with PD. The data is separated by condition and the networks are constructed for each condition. The modules are then identified and their module eigengenes are calculated. The modules and their eigengenes are evaluated and analyzed to identify interesting modules that might reveal biological interesting genes and dysregulated pathways. The genes are then identified within dysregulated pathways that are associated with PD, or by studying topological measures of genes within a module that is thought to contain the causative genes. The flow of this study is illustrated in Figure 16.

4.1 Data

The data needs to be filtered and quality controlled to reduce any noise that might lead to results that are not meaningful. It is also needed to reduce the number of genes and samples to focus on the most varying genes between the conditions. The count matrix is also separated based on the condition of the samples to conduct a differential analysis; PD patients VS healthy persons.

The selected cohort is Park-West(PW). The rows correspond to the samples and the columns represent the genes. The count matrix of this cohort was filtered by identifying "good genes" and "good samples", and any obvious sample outliers were removed. By calculating the median absolute deviation(MAD), the top 10 000 varying genes where identified. The count matrix for PD patients, from now named case, consists of 17 samples. The count matrix for controls consists of 10 samples. Both count matrices contain the same 10 000 genes.

4.2 Over-represented Gene Ontologies

An over-representation analysis of GO terms within a set of genes could give an executive description of biological processes associated with these genes. For this study, the over-representation analysis of GO terms reveals the terms associated with the most varying genes between the controls and the PD patients, which will indicate the processes that vary the most and most likely play a key role in PD. The terms were identified with GO::TermFinder [49] and visualized with REViGO [50].

To identify the GO terms for the analysis, the gene list consisting of the 10 000 top varying genes was analyzed. The p-value cutoff was set to 0.01(default). Among the list of genes, 10 duplicates and 865 identifiers were removed by the GO::TermFinder [49]. In the results, 10 541 terms were found in the biological process domain, but only 647 were displayed due to the p-value cut off. The results are presented in an HTML table format that was imported to REViGO [50] to visualize the terms.

The input field in REViGO for the GO term identifiers and their p-values was pre-filled from GO::TermFinder. The allowed semantic similarity was set to medium(0.7), which was the default. Semantic similarity describes processes that are similar by their annotated genes and their functionalities. The advanced options were left as default too. REViGO summarized the GO terms by hierarchical clustering and visualized similarities and the p-values of the terms in the results.

The scatter-plot and the tree-map(Figure 17) showed the representative pathways for some of the clusters. From both these visualizations "Neuron projection development"-cluster was the biggest cluster, along with "mRNA metabolism", "Modulation of synaptic transmission", "Regulation of catabolism", and "Peptide transport". This indicates that the most varying genes between control

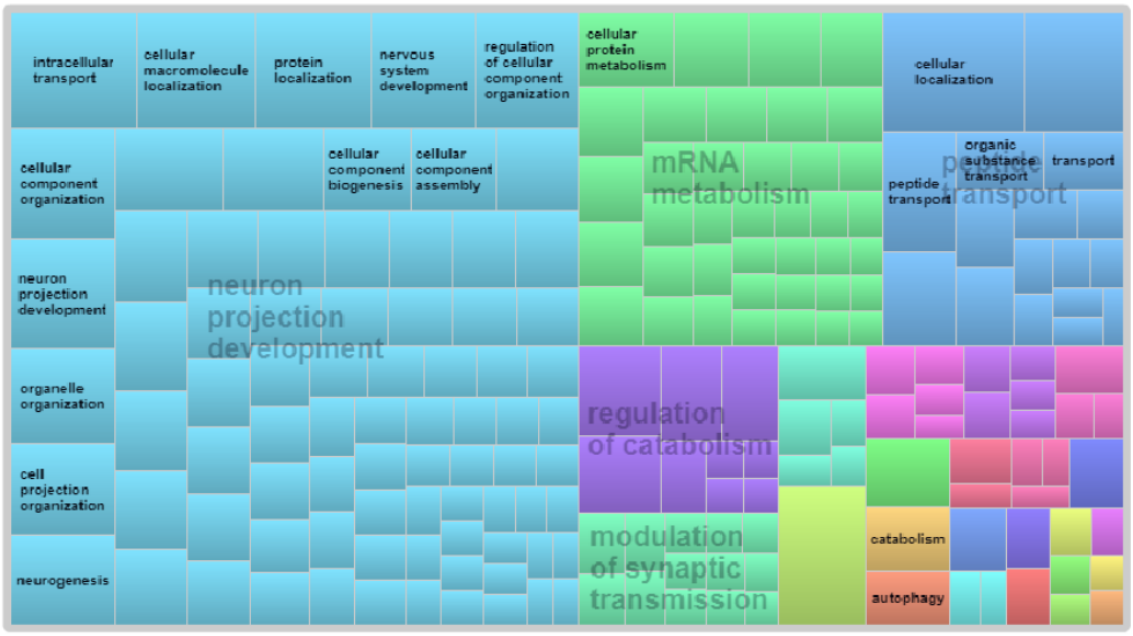

Figure 17: This treemap visualizes the functionally enriched GO terms within the gene set used for this analysis. This visualization indicates the biological processes that are associated with the most varying genes. The block size indicates the p- values of the terms.

and case data set are associated with these biological processes, hence these processes are the most varying processes between controls and PD patients. Terms in the "Neuron project development"-cluster and "Modulation of synaptic transmission"-cluster are recognizable with known molecular dysfunctions associated with PD.

4.3 Constructing the Network

The networks are constructed separately for each condition to do a differential analysis between PD patients and healthy persons. The genes are clustered into modules, which are then merged by the similarity of the gene expression profiles between the modules to avoid too many small modules with high similarity. The step-by-step tutorial of WGCNA was followed with adjustment of some parameters accordingly to each count matrix.

To construct weighted networks for biologically more meaningful networks a soft threshold is set. A set of soft-threshold powers ranging from 1 to 30 is computed for generating the scale-free topology index plot. For both conditions, the plot is very different from each other, as shown in Figure 18. The red line is set at 0.8, and the powers above this line indicate high scale-free topology. From these plots, the soft threshold power 5 is selected for the case network and 24 for the control network.

The soft threshold power is used to modify the adjacency matrix that describes the relations between the gene profiles into a topological overlap matrix(TOM). Then the dissimilarities are calculated by using the TOM, which is then used to create a gene tree for both conditions by

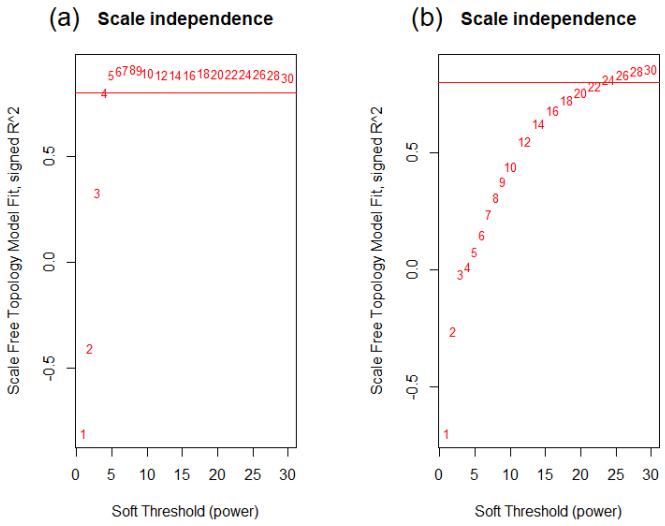

Figure 18: The scale-free topology plots. The soft threshold powers range from 1 to 30, and the red line indicates the scale-free topology fit index is set to 0.8. (a) Case condition; selected power is 5. (b) Control condition; selected power is 24.

hierarchical clustering. The gene trees visualize the networks as dendrograms and are used to identify modules by dynamic tree cut function for branch cutting. The minimum module size is set to 30 and deepsplit-variable is set to 2.

The module eigengenes are visualized in a module eigengene tree by clustering their calculated dissimilarities to identify and merge the close modules. The cut lines for merging the modules are set to 0.1 in the control module eigengene tree and 0.2 in the case module eigengene tree(Figure 19). The module eigengene trees indicate that there are more modules in the control network than in the case network.

The results of the gene trees are visualized in Figure 20, where the color-labels before and after merging the modules are shown for both conditions along the dendrogram. As indicated by the module eigengene trees(Figure 19), the blue, yellow, brown, and turquoise modules are merged in the control network. In the case network black, blue, and turquoise are merged to the black module, the yellow into the grey60 module, and the red into the magenta(Figure 20). The merging causes larger modules because some of the bigger modules that have high similarity are merged. The color labels of the modules indicate that there are some similar modules between the networks that might have some overlapping genes.

The similarities of the gene profiles in the dendrogram can also be visualized by a heatmap plot(Figure 21), where darker shades indicate the modules. The dendrograms and the heatmaps show that both networks have one large module resulting from merging many modules, and some of the color labels appear in both conditions.

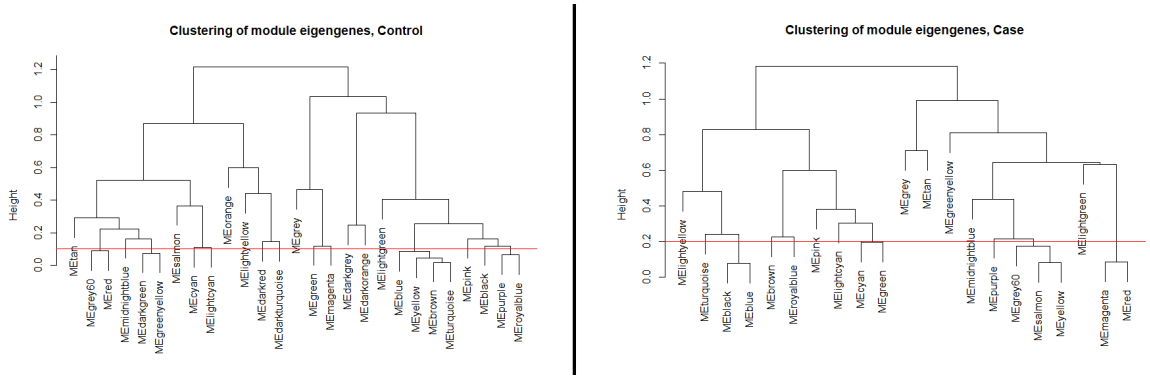

Figure 19: The module eigengene tree for the control network is presented to the left, and the case network to the right. The module eigengene trees indicate that there are more modules in the control network than in the case network.

4.4 Evaluating the Modules

An evaluation of the modules could help to select the more interesting modules, that might contain the dysregulated pathways or the causative genes. The evaluations will indicate similarities between the modules, within a network and between two networks. The module eigengene heatmap shows the relationships among the modules, separately for each condition network. The module correspondence matrix compares the modules in a pairwise manner and finds similarities between them. The module preservation statistics are used to analyze the preservation of the modules from the reference network in the test network. These three methods are used together to evaluate and analyze the modules, for selecting interesting modules for further analysis. In this study, the module labels are used for comparing between case and control, as in the diabetes study [42].

4.4.1 Module Eigengene Heatmap

The adjacency heatmaps of the module eigengenes(Figure 22) show the relationships among the calculated modules for the control and case network separately. This is done separately because the networks are constructed separately and it might be interesting to analyse how the adjacencies differ from control network to case network. The analysis of the module eigengenes and their adjacencies can be useful for finding interesting modules by looking at the changes of similarities from the control to the case network. The color scale ranges from blue to red where red indicates high adjacency and blue indicates low adjacency. The adjacencies describe the correlations between the module eigengenes. The colors representing the rows and the columns are the modules in the network. There are 14 modules in the case network, and 20 modules in the control network.

The color-labels that are only in the control network are blue, green, dark grey, dark orange, dark green, salmon, orange, dark turquoise, and dark red. The color-labels that are only in the case network are green-yellow, royal blue, and brown. The grey module is not in the heatmaps as the genes in the grey module are "leftover" genes, which are the genes that were not assigned to any module.

Comparing the adjacencies of the modules present in both networks indicates that some module adjacencies are different from control to case, and some modules show small differences. When

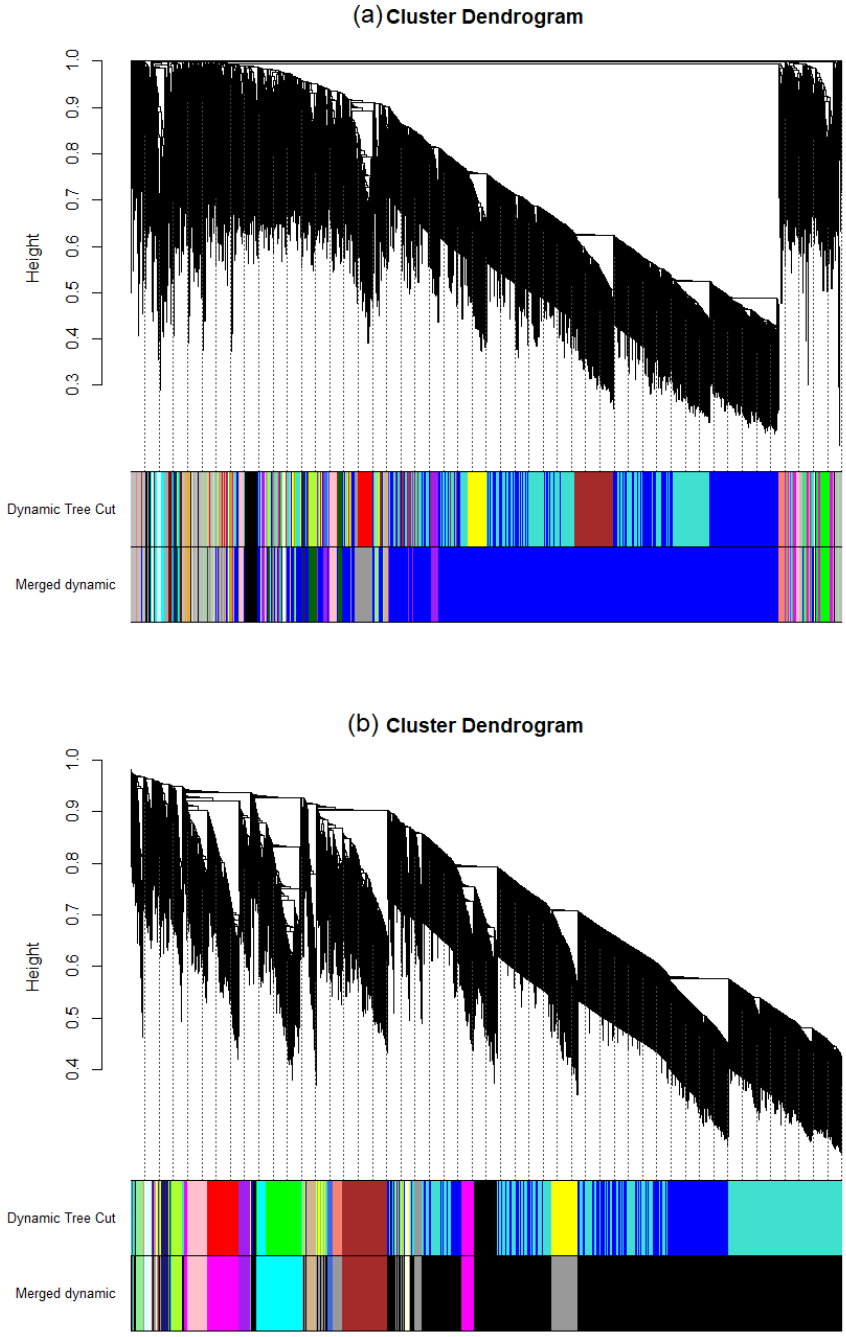

Figure 20: This is the result of network constructions. The dynamic tree cut colors are before merging the modules and the merged dynamic is after merging. (a) The dendrogram for the control condition, and (b) The case dendrogram.

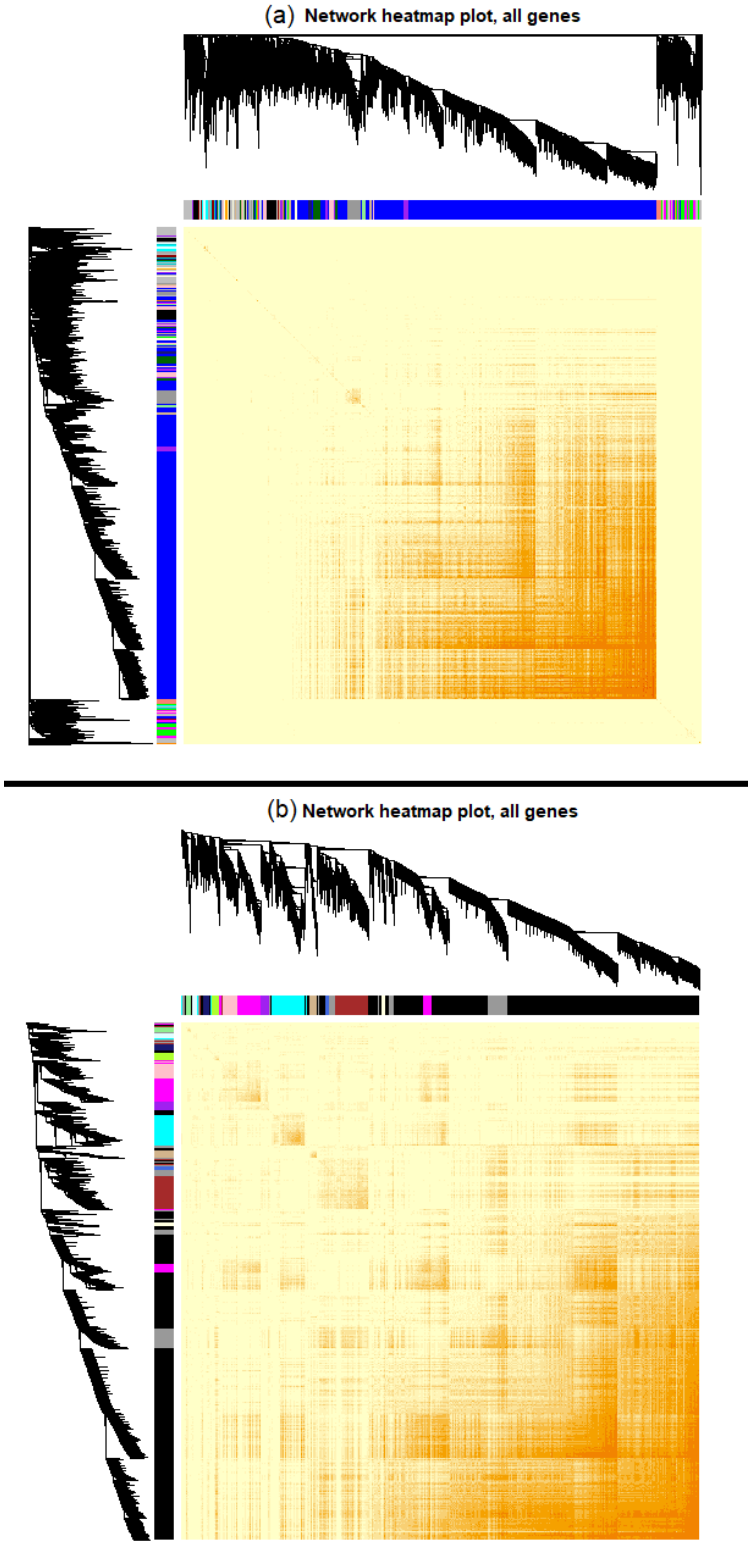

Figure 21: These are the visualizations of the dendrograms in Figure 20, by a heatmap plot. (a) Control network heatmap, where the blue module is the largest, and (b) Case network heatmap, where the black module is the largest.

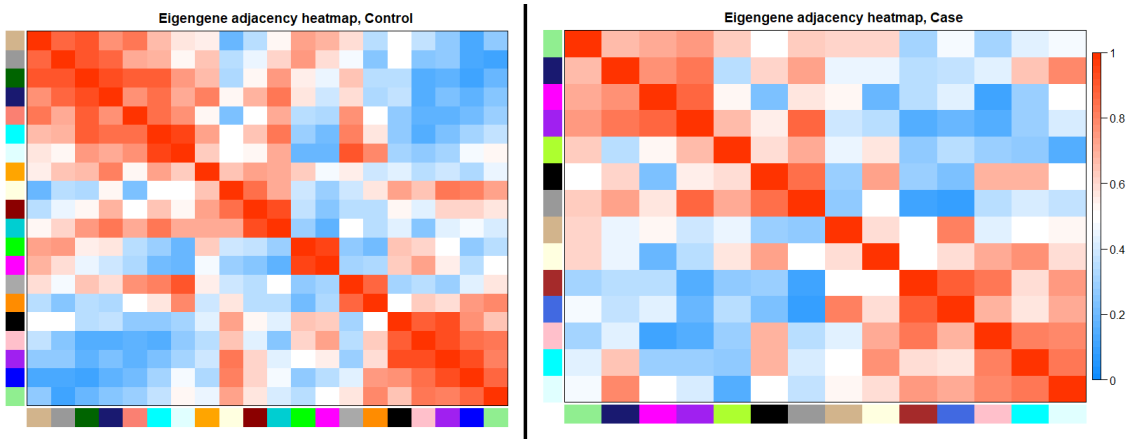

Figure 22: Module eigengene network heatmaps for control(right) and case(left) module eigengenes. The colors presenting the columns and the rows are the modules, and the colors in the matrix ranging from blue to red indicate the similarity measure between the modules within the condition network.

comparing control module heatmap to case module heatmap the pink, black, and tan modules have many modules with a big difference in adjacencies for most of the other modules, and some with smaller differences.

4.4.2 Module Correspondence Matrix

A correspondence matrix compares the modules of two networks and describes the gene count distribution and the similarities of each module. This evaluation allows for comparing the modules by similarity measures calculated by Fisher's exact test. It also allows for comparing the gene distributions in numbers and shows the count of overlapping genes between the modules of both networks.

The module correspondence matrix for case modules vs control modules(Figure 23) is constructed by pairwise comparison. The color-code indicates the similarity measures from white to red, where white indicates low similarity, and red indicates high similarity.

Some modules have high similarity and high overlapping count, for example, the case black and the control blue module. The case black module has overlapping genes with all the modules of the control data set. The color-labels below the case dendrogram in Figure 20 shows that the case black module is a big module after the merging of the modules with similar expression profiles. The merging shows that the case black module includes the case blue module when merging, which might explain the high overlapping count and the high similarity of the case black module with the control blue module.

Many of the color- labels are present in both of the networks, but none of them have markedly high overlapping count nor high similarity measure between the same color modules. They seem to overlap more and have high similarity with other color modules in the other network.

The grey modules contain all the "leftover" genes, the genes that are not assigned to any of the other modules. In the control network, there are 708 genes in the grey module. Only one of

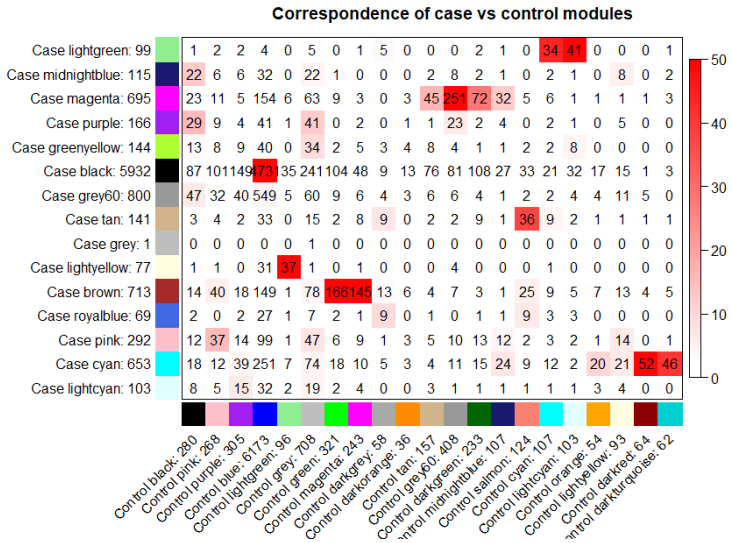

Figure 23: This matrix describes the correspondence of the modules in the case network vs the control network. The color-scale from white to red indicates the similarity measures between the modules, and the counts present the overlapping genes in the modules. The rows represent the color labels of the modules in the case network and the columns represent the control network.

these genes is also in the case grey module as the only gene in this module. The rest of the genes in the control grey module overlaps with all the modules of the case network, where the highest overlapping count is with the case black module. This indicates that the genes in the control grey module are assigned to modules in the case network, except for one gene.

4.4.3 Module Preservation

Module preservation statistics can be used to evaluate the preservation of the modules from a reference network in a test network. This will also indicate which modules that might contain the dysregulated pathways by using the hypotheses presented by Medina *et al.* [42].

By using the module preservation function, module preservation statistics are calculated based on the control network as the reference network and the case network as the test network and their module labels. The $Z_{summary}$ score is plotted against the module size in Figure 24. The blue line(height = 2) presents the lowest preservation and the green line(height = 10) presents the weak preservation. The grey module is not included in this calculation.

The lines indicate the lower levels of preservation, and these are the modules that might contain genes associated with dysregulated pathways of PD. From the $Z_{summary}$ plot(Figure 24), the modules that are below or closer to the blue line are dark orange, orange, dark grey, pink, and black.

The color labels of dark orange, orange, dark grey are not present in the case network after the merging of close modules, as seen in the summaries and plots in previous evaluation sections. The dark orange module shows no similarity with any other module in the case network in the correspondence matrix, which might explain the low preservation. The dark-grey module shows

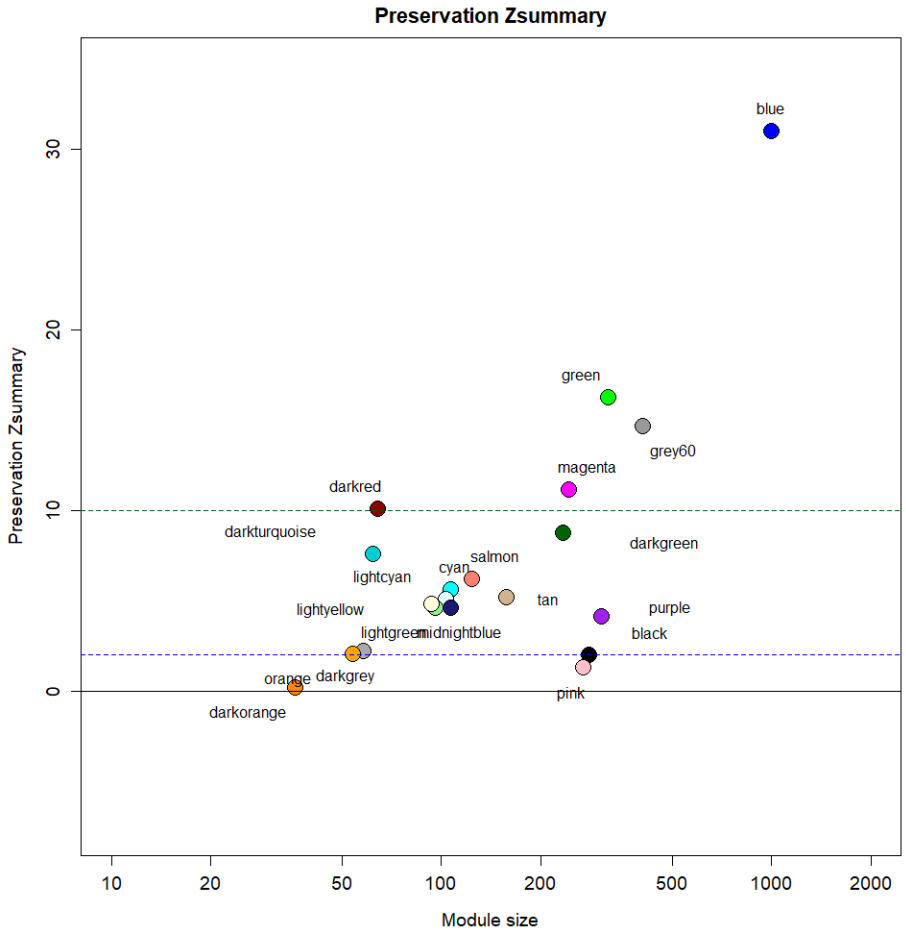

Figure 24: This plot describes the preservation of the control modules in the case network. It shows the preservation statistics by $Z_{summary}$ scores plotted against module size. The modules below the blue line are the lowest preserved modules and the green line indicates weakly preserved modules. Here the control network is the reference network and the case network is the test network.

low similarity in the correspondence matrix with two modules. The orange module shows low similarity with the case cyan module, and have the highest overlap count with this module. The pink and black color labels are present in both networks. The low preservation indicates that these modules are dysregulated in the case network.

The other modules of the control network that are not present in the case network are blue, green, dark green, salmon, dark red, and dark turquoise. The blue module shows the highest preservation in this plot and is also the biggest module by size. More than 4000 of the control blue module overlaps with genes in the case black module, which can be explained by the merging of the case blue module into the case black module. Some modules are only present in the control network that show high preservation. For example, the dark red module and the green module, which both show high similarity with modules of case network.

4.4.4 Interesting Modules

These matrices of module evaluation and the preservation statistics together indicate the interesting modules for further analysis. The interesting modules will most probably contain the dysregulated pathways with causative genes by considering the differences indicated by the evaluations. By the analysis of the eigengene adjacency heatmap, some of the modules with color-labels present in both networks showed difference in the similarities that could be interesting to look at. Further by the three analysis methods the 5 lowest preserved modules that also showed differences in module eigengene network and correspondence matrix are selected; dark grey, dark orange, orange, the pink and black module of the control network, and the pink and black module of the case network.

The pink and the black color label are present in both networks. The pink modules also had high similarity indicated in the correspondence matrix(Figure 23). The pink modules and the black modules also indicated differences in the module eigengene heatmaps.

4.5 ConsensusPathDB

An over-representation analysis of the functionally enriched pathways in the interesting modules may reveal some dysregulated pathways associated with PD. For the functional enrichment analysis of the dysregulated pathways, the online version of ConsensusPathDB. Gene lists from each interesting module were uploaded and KEGG was selected as a pathway source. The p-value cutoff was set to 0.05 and the minimum overlap size with the input list was set to 2.

The analysis gave 108 pathways for case black module, 36 for control black, 10 for case pink, 62 for control pink, 30 for dark grey, 21 for dark orange, and 8 for orange. Among the bigger lists of pathways for the control module gene sets, signaling pathways was a common factor for the lists, also verified by their word clouds. The signaling pathways decreased and the pathways related to diseases increased from control to case modules. For further analysis, the pink and black modules are compared between the conditions, as these color labels are present in both condition networks and the evaluations indicated these as interesting modules. The pink modules had high similarity in the correspondence matrix(Figure 23), and pink and black modules had many differences in the similarities with the other modules described in the module adjacency heatmap(Figure 22).

4.5.1 The Pink Modules

The control pink and case pink module is almost equal by size, where the control pink module contains 268 genes, and the case pink module contains 292 genes. The numbers of pathways were

very different, the control pink list had 62 pathways and the case pink had 10 pathways. There were only three overlapping pathways with some similar genes. In the control module almost 50% are signaling pathways, but there is only one signaling pathway in the case pink output list.

In the pink control output list "Dopaminergic synapse" was listed with 5 candidates. This indicates that this pathway is associated with the most varying genes within PD. The presence of this pathway can also show that the dysfunction is not directly in the dopaminergic synapse, but in an earlier or later step of the signal transmitting. In control pink, the genes that overlapped with the dopaminergic pathway are *PRKCA, AKT1, MAPK11, CAMK2A*, and *GNB2*. All of these genes are related to signaling by being directly involved in cellular signaling pathways or involved in the integration of biochemical signals for a variety of cellular processes [57]. Out of these genes, only *AKT1* is present in the case pink gene set, but because of the minimum overlap size set to 2, it will not give a match with the "Dopaminergic synapse"-pathway.

On the other side "Dopaminergic synapse" is listed in the black case output list with 60 candidates, which is the opposite patient group. According to the correspondence matrix, the control pink module overlaps with the case black module with 101 genes, but there are no overlapping genes in the candidate list of the pathway between control pink and case black output-lists.

4.5.2 The Black Modules

The resulting list of control black module contained 36 pathways, and the case black module contained 108 pathways. The different sizes of the lists are because of the different sizes of the gene sets. The control black module has 280 genes and the case black module has 5932 genes. The case black module contains a larger set of genes, hence the pathways also contain many genes per pathway.

These two modules have a total of 21 common pathways, where some of them had overlapping genes. Many of the common pathways are signaling pathways. In the control black module output list approximately 36% of the pathways are signaling pathways, and in the case black module it is around 25%. This shows that the signaling pathways might be dysregulated from control to case.

Another thing to compare between these lists is the disease-associated pathways. In the control output-list, it is only 8% of the pathways that are associated with a disease, but in the case output-list, it is approximately 14.8%. Among these disease pathways in the case black module, Parkinson's disease is one of them along with Huntington's disease and Alzheimer's disease. 99 of 5932 genes match with the gene set representing Parkinson's disease in KEGG. The recognizable genes are *PARK7*(*DJ-1*), *LRRK2*, *SNCA*, and *UCHL1*. These are described in Table 1 as early-onset PD(EOPD) with few cases and good responses to levodopa treatment, except *LRRK2* which is found in later-onset PD(Classical PD).

Another interesting pathway identified within the case black module was the "Ubiquitin mediated proteolysis" pathway, which has an important role in cellular processes with the functionality of protein ubiquitination [40]. This system regulates the presence of reactive oxygen species to protect the cell for extra oxidative stress and mitochondrial damage.

4.6 ClueGo

ClueGo is a plug-in tool in Cytoscape that can be used for functional analysis of pathways and how they interact by common genes with network visualization of the pathways. The database used for this tool is also KEGG so that the results can verify the findings of the analysis in ConsensusPathDB, and compare the pathways found in both networks.

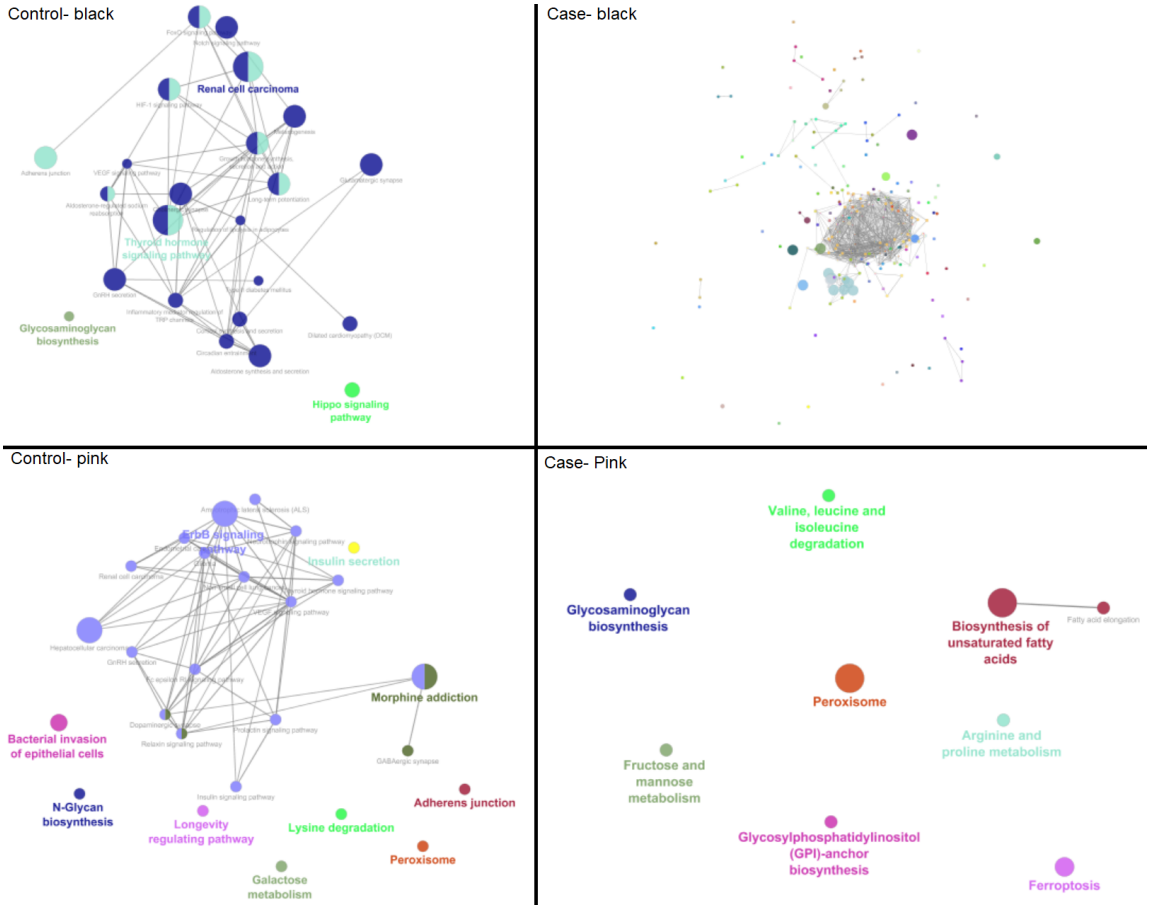

Figure 25: These are the pathway networks from ClueGO analysis for each set of nodes from the module networks. The colors indicate functional groups, and the edges are weighted by a Kappa score. Some nodes have multiple colors, which indicate that this node is a part of multiple functional groups. The levels of detail are different from module to module.

ClueGo creates a network of the pathways, and the layout tells which pathways are functionally similar and how similar they are. The modules used in this tool are pink and black modules from both case and control networks, the same that were used for module network topology analysis. This means that there are fewer genes than in the ConsensusPathDB analysis, because of the filtering done when making gene sets for network topology analysis. These modules were also selected for this analysis as these were the most interesting modules in the ConsensusPathDB analysis.

Figure 25 shows the results of the ClueGo analysis. They vary in how detailed they are, and also in how similar the pathways are. The numbers of pathways and functional groups are also different from module to module. To present the network in this figure, some of the nodes that were far away from the others were moved closer to get a better overview of all the pathways in one square. The distance in the original network tells how similar the pathways are.

4.6.1 Pink Modules

In the control pink module network, there were found 10 functional groups. The toggle was set right above medium to get the most important functional groups and pathway, as a more detailed one was more complicated to analyze. The largest group is the "ErbB signaling" pathway, where all the signaling pathways are connected. ErbB signaling pathway is defined as a signaling pathway that regulates different biological processes such as regulating cell survival [40].

The edge colors representing the Kappa score indicated strong connections between signaling pathways and their functional groups. These signaling groups are associated with functions such as signal transductions which are important for cell survival. One interesting pathway in this module is "Dopaminergic synapse", also found in the analysis of ConsensusPathDB. Dopaminergic synapse is an important synapse pathway for the neurotransmitter dopamine(DA) in the brain. DA controls different functions such as motor signaling and more in the central nervous system [40]. Defects in this pathway may lead to a lack of dopaminergic neurons which causes many of the PD symptoms. The connectivity to other signaling pathways explains the importance of signals transmitted by neurons in dopaminergic synapse because this pathway also activates several other signaling pathways.

Lastly, the ClueGO analysis was done for the case pink module. The toggle was set in the middle of medium and detailed as this network did not have many pathways. The more detailed network gave more pathways, but no more connections than shown in Figure 25. In this network of pathways, there were 8 functional groups, and only one group had two pathways connected as visualized in Figure 25. This case module network gave no match with signaling pathways.

4.6.2 Black Modules

The control black module contains significantly fewer genes than the case black module, which resulted in fewer pathways as in the ConsensusPathDB analysis. The 60 nodes of the control black pathway network are very similar, and only 2 of them are not linked to any other pathway. The control black module resulted in 4 functional groups where 2 of them are strongly connected and cover most of the pathways. The two functional groups that are connected are the "Thyroid hormone signaling" pathway and "Renal cell carcinoma" pathway, and they share some pathways in between them. The Thyroid hormone signaling pathway is central of activating many other signaling pathways, essential for many biological processes [40], which also explains the connections to the other signaling pathways.

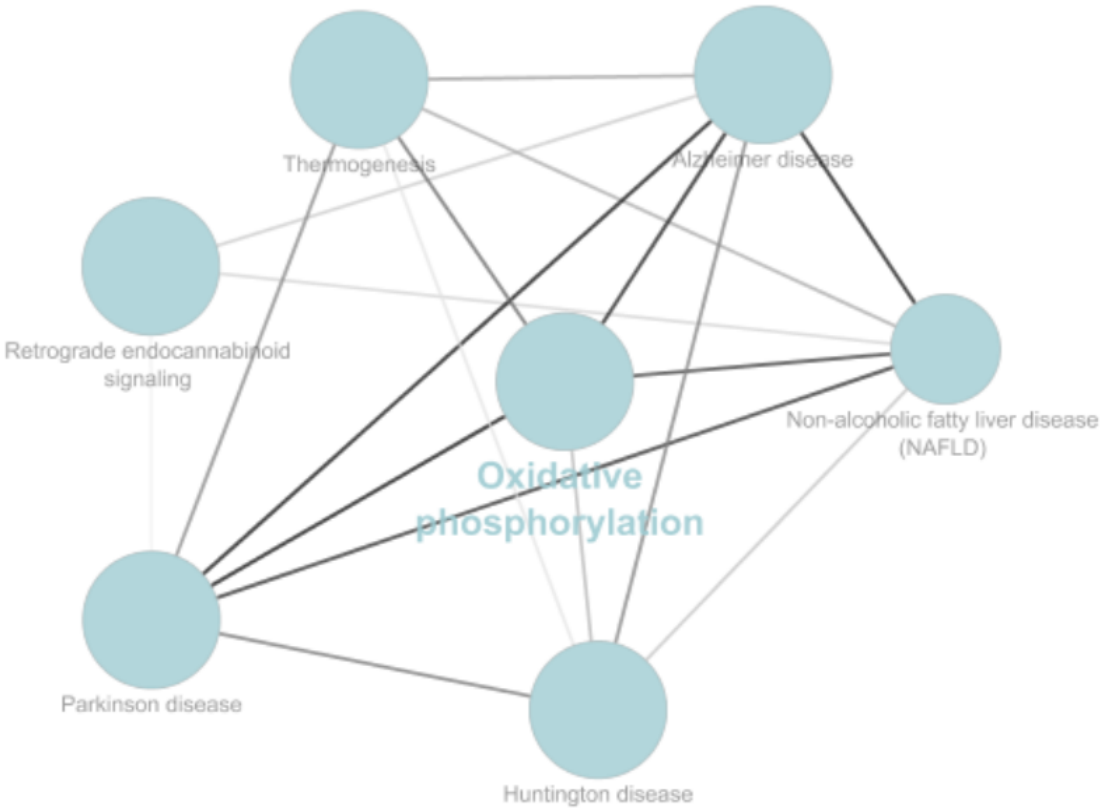

Figure 26: This sub-network is a part of the case black ClueGO network in Figure 25. These nodes belong to the functional group of oxidative phosphorylation as the representative pathway, and Parkinson's disease is a part of this sub-network. The edges are colored by the Kappa score, where the darker shade of grey indicates a higher score.

The case black module is the largest module of the case network, and in ConsensusPathDB the output contained a high number of pathways. For the module network analysis, the number of genes was reduced to 533, but it still resulted in many pathways. To avoid too many pathways the toggle button was right below medium for this gene set, which still resulted in many pathways. The pathways were divided into 140 functional groups, where many of the groups only contained one pathway. Many of the pathways are similar and close together in the "hairball" in the middle, but the other nodes far away from each other indicate a high variation of pathways within this module. In the analysis of the ConsensusPathDB, there were some interesting pathways such as "Parkinson's disease" and "Ubiquitin mediated proteolysis". Ubiquitin mediated proteolysis pathway has high p-value in the ClueGO network of this module, but it is not connected with any other pathway. In this network, the Parkinson's disease pathway belongs to the functional group of "Oxidative phosphorylation" where all the nodes have a high p-value, hence the larger size nodes.

The functional group oxidative phosphorylation contains three neurodegenerative diseases, which

Genes	Association/Functionality
ADCY5, ADORA2A, PRKACA, PRKACB	Signaling compounds and pathways
MT-ND1, MT-ND2, MT-ND3, MT-ND4, MT-ND4L, MT-ND5, MT-ND6 S*LC25A4, SLC25A5, SLC25A6* (Translocates ADP/ATP)	Respiratory chain
LRRK2, *PARK7*, *PINK1* (Section 1.1.2) *SEPT5*	Parkinson Disease
UBE2J1, *UBE2L3*	Modification of proteins with ubiquitin
VDAC1, VDAC2, VDAC3	Mitochondrial integral membrane protein
PPIF	Involved in ATP synthase activity and may assist in protein folding

Table 2: This table shows the genes that are in Parkinson's disease pathway and not in Alzheimer's disease pathway from ConsensusPathDB analysis, with their associations and functionalities. Information is from genecards [57].

explains the similarity between Alzheimer's Disease, PD, and Huntington's disease. In Figure 26 the functional group of Oxidative phosphorylation is extracted with the adjacent nodes. None of these pathway nodes are connected to other nodes in the remaining network. For this closer look, the edge color is changed to a gradient greyscale indicating the kappa score, where darker shades of grey represent a higher Kappa score.

Oxidative phosphorylation is a biological process that phosphorylates ADP to ATP which is the fuel of the cell, which takes place in the mitochondria [58]. The oxidative phosphorylation happens through the respiratory chain which consists of five complexes, where dysfunction in complex I is found to be associated with PD and is also a reason for increased production of ROS [59].

The stronger connectivity from Parkinson's disease to Alzheimer's disease, than to Huntington's disease, confirms that these two diseases are very similar in dysfunctionalities. By looking at the genes that matched with Parkinson's disease pathway in the list from ConsensusPathDB compared to Alzheimer's Disease, the genes that are different between these disease-associated pathways can be identified. As the Kappa score tells there are many similar genes in both of the sets, but there are some genes that are only in either of them which can be a major difference to tell the diseases apart from each other.

The genes are listed in Table 2 and categorized based on common association and functionalities. The largest group is the genes associated with the respiratory chain, which explains the strong connectivity to the Oxidative phosphorylation pathway. Along with *LRRK2*, *PARK7*, and *PINK1* which is explained to be associated with PD in section 1.1.2, *SEPT5* is also associated with PD [57]. Over-expression of *SEPT5* is associated with dopamine-dependent neurotoxicity, and degradation of this gene may lead to early-onset PD(EOPD) [57].

In this set of genes, two genes are associated with the modification of proteins with ubiquitin, which has an important role in decreasing oxidative stress and preventing mitochondrial damage. The presence of this pathway in the case black module network, and that two genes of Parkinson's disease pathway are associated with ubiquitination may indicate that there is a higher level of oxidative stress by reactive oxygen species in PD genes compared to Alzheimer's disease genes.

4.7 Module Network-Cytoscape

The network analyzer tool in Cytoscape provides a statistical analysis that describes the topological properties of the network for each module network. The statistical analysis calculates many topological measurements, where the betweenness centrality and the node degree is kept in the network tables for further analysis.

In the module networks, the nodes represent the genes and the edges are weighted by the adjacency measures from the topological overlap matrix. The nodes and the edges are filtered by selecting the top-ranked genes by their soft connectivity measures, and by setting a weight threshold for each module network.

The modules that are visualized in Cytoscape are the control and case modules of black and pink(Figure 27). The case black module is the largest module with more than 5000 genes, even after filtering. For this module, the 1000 top-ranked genes are selected before setting the weight threshold by calculating soft connectivity measures. The case black module had a threshold on 0.57 with 533 genes and 3529 edges, which will say that the weight threshold left approximately 50% of the genes disconnected to the resulting network. For the control black module, the weight threshold is set to 0.06 which resulted in a network of 275 genes and 2985 edges. The pink modules have the threshold 0.1 for the control module and 0.2 for the case module, where the control module contains 235 genes and 4007 edges, and the case module 262 nodes and 3988 edges.

In the visualization of the module networks in Figure 27, the network node's color is a gradient with the node degree from the network topology analysis. The color scale is from light-yellow to orange to red, where yellow indicated a low degree, orange medium, and darker red high degree. The large gene sets in each module formed "hairballs" in the network visualization, making the table resulted from the statistical analysis more useful for further analysis.

4.7.1 Identifying Genes by Analyzing Betweenness Centrality Measures

Betweenness centrality(BC) measure indicates biological information transfer, and how important each gene is in a biological network [42]. The overlapping genes between the control and case module networks can be used to analyze and compare the importance of the genes in the networks. Then the roles of the genes can be associated with functions causing PD.

The genes with the highest betweenness centrality measure in each module network will have a central role in the module. In the control pink module, the *PITPNM3* gene has a BC measure equal to 1.0 and in the case pink module the *CYB5R1* has the highest BC measure at 0.0999, and are the genes with the highest BC measure in these modules. The *PITPNM3* gene encodes membrane-associated transfer proteins and the *CYB5R1* gene is related to oxidoreductase activity [57]. In the black control module, the *IQSEC2* has a BC measure 0.076 and in the case black module the *SEC63* gene has the highest BC measure that is 0.2799. The *IQSEC2* gene may play a role in the cytoskeletal and synaptic organization, where mutations of this gene have been associated with cognitive disability [57]. The *SEC63* gene from the case black module is associated with the unfolded protein response pathway and is a central component of the protein translocation apparatus of the endoplasmic reticulum(ER) membrane [57].

Overlapping genes between the module networks enables a comparison of the genes by their BC measure between the modules of the same module color. In the pink module networks, there are 30 out of 37 overlapping genes after setting the weight threshold, and in the black module networks, there are no overlapping genes left after the filtering. The black modules are very different in size, and approximately 50% of the genes are filtered out in the case black modules. No overlapping

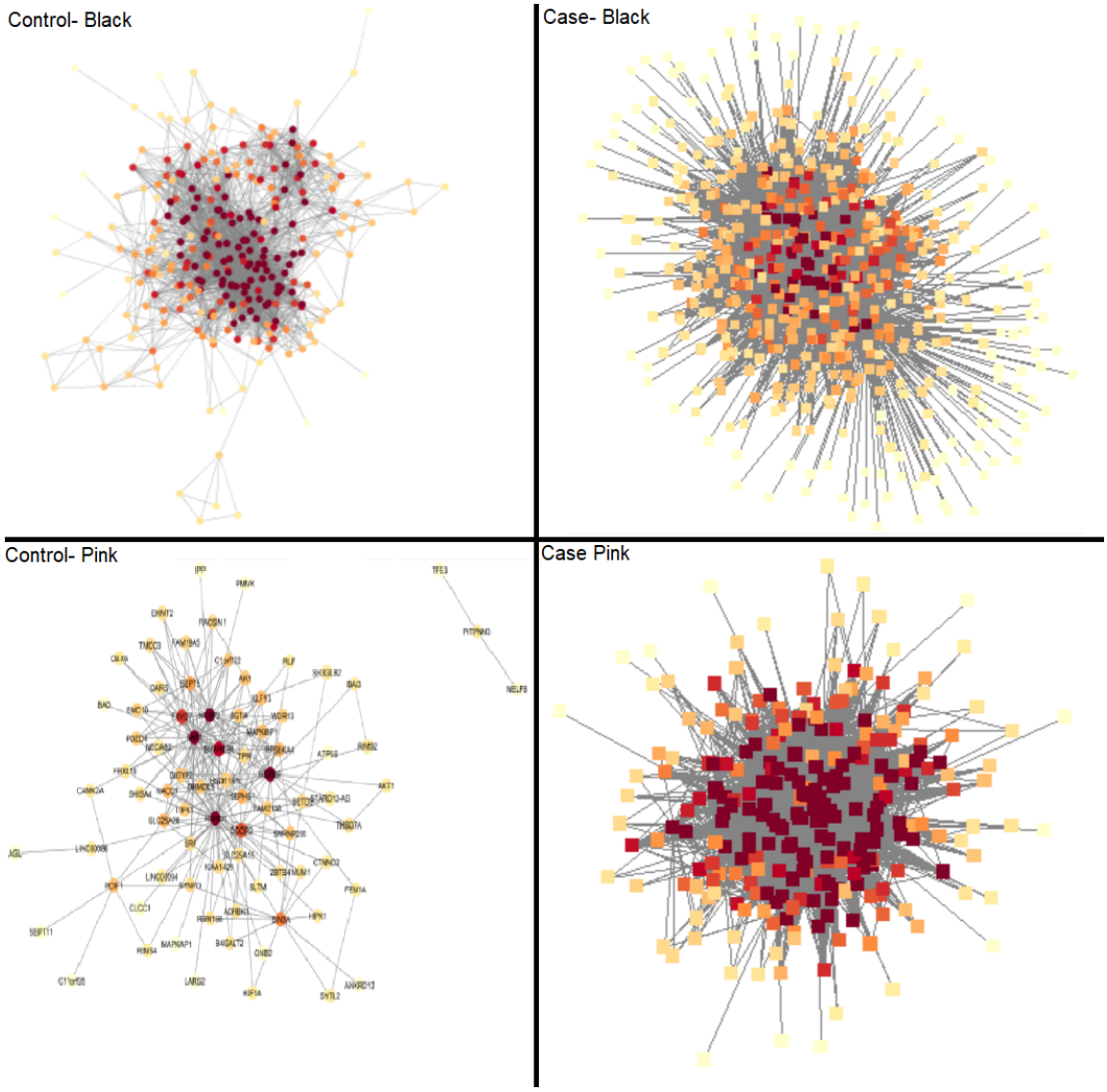

Figure 27: Visualization of the module networks in Cytoscape. The nodes are colored by their degree from yellow to dark red. Yellow indicates a low degree and darker red indicates a high degree.

genes in the black modules indicate that the 87 overlapping genes were not among the genes with the highest soft connectivity measure in the case black module.

By looking at the overlapping genes in the pink module and their betweenness centrality(BC) measure from control to case, the similarities of the BC measures can be compared. The genes with very similar BC measures are *SNX8, TPR,* and *AKT1*. The genes that were very different are *PCM1, AK1, NCBP2, DARS, IVD, WASF3, SIN3A, NELFB, NLGN3, IPP*, and *Clorf122*, and they all show lower BC measure in the case module network than in control module network. This indicates that these genes that were different have less important roles in the pink module of PD patients. Out of these, *AKT1, AK1, CSPG5, IVD*, and *NLGN3* have some interesting functionalities that could be related to the known symptoms and dysfunctionalities of PD.

AKT1 is a protein-coding gene that is related to many signaling pathways and cell survival associations [57]. This describes the findings with ConsensusPathDB where *AKT1* is related to many of the pathways, especially signaling pathways. This gene regulates cell survival by phosphorylation of other genes [57]. *AKT1* has a betweenness centrality equal to 0 in the case and control module network of the pink modules.

AK1 is found to be highly expressed in skeletal muscle and brain and certain mutations in this gene are associated with a rare genetic disorder that can destroy red blood cells [57]. *AK1* is also associated with metabolic pathways.

CSPG5 encodes a protein that may function as a neural growth factor that is essential for the regulation of growth, maintenance, and survival of neurons [57]. Mutations or lack of this gene may cause a higher rate of neuronal loss, which might explain the lower BC measure in the case module.

A lack of *IVD* results in an accumulation of an acid that is toxic to the central nervous system [57]. The BC measure for this gene is 0 in the case network and higher in the control network. This may explain more toxicity to the central nervous system with PD patients, as this gene has a more important role in the control module.

NLGN3 encodes neuronal cell surface proteins that play a role in synapse function and synaptic signal transmission, which may have a role in the formation and the remodeling of the central nervous system synapses. The lower BC measure in the case module may indicate dysfunctions of the synapses.

4.8 Interesting Genes

The analysis methods of ConsensusPathDB, network topology analysis, and the pathways visualized by ClueGO indicated some interesting genes in the selected modules, by having functionalities that could be associated with PD. Table 3 presents a summary of all the genes mentioned as interesting genes in the analysis of the modules. The genes are described with what module they were found in, their functionality/association, and what analysis they were found in. The table is sorted by the module they are found in, and for the pink module, some genes are in both condition modules.

4.8.1 Pink Module

CYB5R1 is the only gene from the pink case module without being in the control module also, and this gene was found when analyzing BC measure as the highest BC measure in this module. This gene is associated with oxidoreductase activity [57].

Many of the genes from the pink control module are associated with the dopaminergic synapse pathway, identified with ConsensusPathDB. *AKT1* gene is present in the pink module for both condition networks and is related to many signaling pathways as well as cell survival associations

Gene	**Module/Condition**	**Functionality/Association**	**Analysis section**
PARK7(DJ-1)	Black/Case	Parkinson's Disease	ConsensusPathDB
LRRK2	Black/Case	Parkinson's Disease	ConsensusPathDB
SNCA	Black/Case	Parkinson's Disease	ConsensusPathDB
UCHL1	Black/Case	Parkinson's Disease	ConsensusPathDB
SEC63	Black/Case	Responsible for unfolded protein response + protein translocation	Betweenness Centrality
SEPT5	Black/Case	Parkinson's Disease	Cytoscape-ClueGO
UBE2J1	Black/Case	Modification of proteins with ubiquitin	Cytoscape-ClueGO
UBE2L3	Black/Case	Modification of proteins with ubiquitin	Cytoscape-ClueGO
IQSEC2	Black/Control	Cytoskeletal + Synaptic organization	Betweenness centrality
CYB5R1	Pink/Case	Oxidoreductase activity	Betweenness centrality
PRKCA	Pink/Control	Dopaminergic Synapse	ConsensusPathDB
MAPK11	Pink/Control	Dopaminergic Synapse	ConsensusPathDB
CAMK2A	Pink/Control	Dopaminergic Synapse	ConsensusPathDB
GNB2	Pink/Control	Dopaminergic Synapse	ConsensusPathDB
PITPNM3	Pink/Control	Membrane transfer proteins	Betweenness centrality
AKT1	Pink/Control+Case	Dopaminergic Synapse	ConsensusPathDB + Betweenness Centrality
AK1	Pink/Control+Case	Metabolic pathways	Betweenness Centrality
CSPG5	Pink/Control+Case	Neural growth	Betweenness centrality
IVD	Pink/Control+Case	Degradation + Metabolic pathways	Betweenness centrality
NLGN3	Pink/Control+Case	Synaptic signal transmission	Betweenness centrality

Table 3: This table summarizes the genes identified in different analyses of the condition networks. The genes are sorted by what condition module they were found in. Their functionality and the analysis they were found in is also described. Many of the genes are found within pathways associated with PD.

[57]. *PRKCA* is a gene that encodes for a protein that plays a role in many different cellular processes [57]. *MAPK11* encodes for a protein kinase that is central in the integration of biochemical signals for many different cellular processes [57]. *CAMK2A* is crucial for encoding protein involved in calcium signaling [57]. *GNB2* is also associated with signaling and involved in signal-transducing receptors. The last gene that was only found in the pink control module is the *PITPNM3* gene that is associated with membrane proteins and had the highest BC measure in this module.

The rest of the genes in Table 3 are also found in BC measure analysis where these were among the overlapping genes. These genes were found to have a higher BC measure in the control module network than in the case module network. *AK1* is described as a highly expressed gene in skeletal muscles and the brain [57]. *CSPG5* is essential for neuron growth, maintenance, and survival [57]. A lower expression of this gene might indicate less growth and differentiation of neurons, and a higher rate of neuron death. Lack of *IVD* is thought to result in a toxic environment in the central nervous system [57], and lower expression of this gene within the case patients might describe more toxicity in the central nervous system of the case patients. *NLGN3* is essential for synapse function and signal transmission which may have a role in the formation and remodeling of the central nervous system synapses [57].

4.8.2 Black Module

The first genes are the ones in the black case module network, where most of them that were found to be interesting are associated with "Parkinson's disease" pathway and some of them are described in section 1.1.2.

PARK7, *LRRK2*, *UCHL1*, and *SNCA* are described in Table 1 as associated with PD. *PARK7* gene encodes proteins that are thought to protect the cell against oxidative stress and avoid cell death, and a defect in this gene causes the autosomal-recessive EOPD [57]. *LRRK2* is associated with the outer membrane and signaling, and *UCHL1* is especially expressed in neurons [57]. *SNCA*(Synuclein Alpha) is highly expressed in the brain and is thought to be involved in presynaptic signaling and membrane trafficking [57]. This gene is also associated with the respiratory chain and ATP synthesis. These genes were found when analyzing pathways that were identified with ConsensusPathDB.

A gene that was not mentioned in Table 1 is *SEPT5*, also called *SEPTIN5*, which is associated with neurotoxicity that may lead to EOPD. This gene was found when analyzing pathways identified with ClueGO. The next gene listed from the case black module is *SEC63*, which is responsible for unfolded protein response and associated with protein translocation, and was identified when analyzing betweenness centrality with the highest measure in the case black module. The last two genes *UBE2J1* and *UBE2L3* are associated with ubiquitination, which is an important process for reducing oxidative stress by removing reactive oxygen species that increase oxidative stress. These genes were also identified when analyzing the PD pathway in ClueGO results.

The only gene from the black control module was *IQSEC2* from BC measure analysis, where this gene had the highest BC measure in this module. This gene may play a role in the cytoskeletal and synaptic organization [57], which can be essential for neuron transmitting and signals in the nervous system. Mutations of this gene are associated with a cognitive disability, which can cause symptoms of PD.

Step	Description	Function	Parameters
Data filtering and selection	Good genes and samples	*goodSamplesGenes*	count matrix, **verbose**
	Sample clustering	*hclust*	distance matrix, **method**
		cutreeStatic	Sample tree, **cutheight**, **minSize**
	Top varying genes	*mad*	count matrix
Network construction	Soft threshold values	*pickSoftThreshold*	count matrix, **cutheight**, **verbose**
	Adjacency matrix	*adjacency*	count matrix, **power**
	Gene tree	*hclust*	distance matrix, **method**
	Module identification	*cutreeDynamic*	dendrogram, distance matrix, **deepSplit**, **minClusterSize**
	Module eigengenes	*moduleEigengenes*	count matrix, colorlabels
	Module eigengene tree	*hclust*	distance matrix, **method**
	Merging modules	*mergeCloseModules*	count matrix, color labels, **cutHeight**, **verbose**

Table 4: This table summarizes the automatic functions of WGCNA used in this study. The highlighted parameters are the parameters that can be tuned.

5 Discussion

Parkinson's disease is one of the most common neurodegenerative diseases with symptoms that makes daily chores difficult. There are several studies ongoing to finding causative genes in PD patients to identify genetic differences that differ PD from other similar diseases and differences between healthy and PD patients. The biological functions that are believed to cause this disease are partially identified, and with that knowledge, genes that play different important roles in those biological processes can be identified. This study approaches dysregulated pathways and causative genes by a differential network analysis based on the co-expressed genes in PD patients vs healthy persons. Modules are helpful to analyze the large data set, and evaluations of these help to identify interesting modules. Further on, the interesting modules were analyzed by different tools and methods which led to interesting pathways and genes that could be associated with PD. Figure 16 describes a flowchart of this study.

5.1 Functions of Network Construction

In the data filtering and network construction, there are many automatic functions. Especially the network construction step consists of many tuned parameters. Tuning the parameters is a major challenge, as there might not be good descriptions of how to select appropriate parameters. Some of the parameters were default and some were changed by testing different values where the parameters were selected when the output was close to what the tutorial had. Another challenge of automatic functions is the reproducibility, because of all the tuned parameters. All the functions of data filtering and network constructions are summarized in Table 4, where the parameters that are tuned are highlighted. This was also a challenge with the tools when setting different cutting parameters.

These large data sets can be challenging to analyze in detail, therefore the network construction and WGCNA methodology in identifying modules were helpful for a detailed overview. The modules were created with different automatic functions, where some parameters were also tuned by testing. These parameters were tuned considering the numbers of the modules the tuning resulted in, but it was challenging to get the same number of modules for both the networks. The different numbers of modules, the different color labels, and the different module sizes in the networks made it challenging to compare the modules in further analysis.

Block-wise network construction is a method to consider for network construction with large data set. As the method name suggests, this method divides the data set in blocks of set size and constructs a network for each block, and is described to be faster. The challenges of this method were to visualize the networks by heatmaps as the TOMs are saved in separate files and to compare and analyze the modules with more than 20 networks. In the tutorial [28], the network that represented the data set the best was selected. The step by step network construction was easier considering the steps after network construction, and as it was feasible to run on the computer it was not necessary to perform to blockwise network construction.

5.2 Evaluating the Modules

The differences in similarities between the modules, when comparing the eigengene heatmaps of the case and control network indicate that the overlapping genes in some of the modules are differentially expressed. It could also be because the modules of the same color label consist of different genes, also indicated by the correspondence matrix.

The correspondence matrix showed that the modules that were in both networks did not always contain the same genes, which could be explained by different clustering and merging of the modules, and that the genes are differentially expressed in both networks as these genes are the most varying genes of the original data set. Every module from both networks shared some similarity with at least one module from the other network, but not many shared a strong similarity or any similarity with the same color labeled module in the other network if the color label was present in both networks.

In the module preservation statistics analysis it was expected that the modules that were not in the case network would show low preservation, as this statistic shows how well the modules are reproduced in the test network. Surprisingly some of them were in the middle and some even showed high preservation. The modules with low preservation were verified by the other evaluation methods as well.

The tutorials of module evaluation methods excluded the grey module(the ”leftover” genes) from preservation statistic analysis and the module eigengene heatmap, but the difference in how many genes that were in the control grey module vs the case module indicated that this module could also be interesting to include in the preservation analysis.

5.3 Tools

Based on the evaluations, interesting modules were selected for further analysis of pathways and network topology. The different tools are summarized in Table 5. The gene sets were the genes from each module selected as interesting modules, and KEGG was selected as the database for the pathways for both methods.

ConsensusPathDB	Cytoscape-BC	Cytoscape-ClueGO
Online tool	Software	Plug-in app to software
A list of matching pathways for each module	Access to network topology measures, and module network visualization	Visualization of the pathways and their interactions for each module
Used to identify dysregulated pathways with the gene-sets of each interesting module	Used to visualize the module networks and to look at the network topology measure betweenness centrality measure	Used to validate pathways found with ConsensusPathDB, and to look at some interacting pathways of the dysregulated pathways

Table 5: Summaries of the tools and what they were used for in this study.

5.3.1 ConsensusPathDB

In ConsensusPathDB the threshold and the minimum overlap size were the same for all the gene sets, but the size of the gene sets varied and resulted in different output list sizes. Maybe different threshold values would have given more equal list sizes, but for the comparing between the modules, without having to consider the parameters, all the thresholds were equal.

The output list could be downloaded, which was found to be more informative, as the downloaded file was presented in a more informative way. The downloaded file contained all the genes that matched with one pathway, and the online list only showed how many genes that matched with the pathway. The genes listed in the downloaded file were very useful when identifying interesting genes within the dysregulated pathways.

5.3.2 ClueGO

The ClueGO plug-in verified some of the pathways listed in the output of ConsensusPathDB, but there were also some pathways left out and some new pathways identified. The functional groups in ClueGO made it easier to look for interesting pathways and pathways similar to them, by using the pathways from the analysis with ConsensusPathDB.

The kappa score also made it easier to identify interesting pathways for further analysis and to extract pathways that were similar within functional groups. This score was used to identify pathway similarities within the functional group oxidative phosphorylation in Figure 26, which contained PD, Alzheimer's, and Huntington's disease pathways. The similarities between these diseases, and that they are hard to tell apart when diagnosing elderly patients are discussed in many studies.

The different networks of the pathways for each module have different levels of details. Some modules needed higher levels of details to get more connections between the pathways, and some modules needed lower levels of details to only show the most important connections between the different pathways. This is also another tuned parameter to consider, which makes it more difficult to reproduce the results.

5.3.3 Cytoscape

Cytoscape is a software for visualizing the networks and it provides extra tools for analyzing the networks. The nodes and edges of the module networks were filtered in different steps of ranking and filtering, considering the numbers of nodes(genes) and edges(connectivity) to be as equal as possible for easier comparison. The thresholds were different for each module network, as the weights varied from module to module, and the number of nodes and edges were considered. The filtering of nodes and edges made it easier to analyze the networks with only the strongest connections among the genes in the modules and between the conditions, but it also decreased the number of overlapping genes between the condition modules. The overlapping genes were useful to look at how the same gene acted differently between healthy and PD patients by comparing modules based on their color label. Even though the filtering made it easier, it was still a challenge with a large number of nodes and edges per module network.

5.4 Evaluating Results

The network construction, module identification, evaluation, and the analysis resulted in interesting dysregulated pathways with genes associated with dysfunctions of PD. One should consider how reliable the results are, and if some selecting and filtering can affect the results. First, the tuned parameters and the evaluation which resulted in selecting only five, later on only two modules. Next, interesting pathways were selected based on recognizable functionalities. Finally, interesting genes were selected by network topology and in the interesting pathways. This study resulted in a list of 20 genes out of approximately 50 000 genes.

5.4.1 Modules

Different tuning of the parameters could have resulted in more modules, and maybe the modules could be related to more specific functionalities by containing fewer genes per module. The modules in this study included a variety of pathways which made it difficult to pin a module to a functional group. This variation might also be caused by the merging of the modules which made some modules larger, and one or two modules containing a large number of the total amount of the genes. The variations within the modules were also indicated by the visualization in ClueGO analysis, where many functional groups were indicating a high variation of functionalities.

In the dendrograms created for each condition, the genes have different correlation measures in the different conditions which results in different assignments of genes to modules. When the genes are clustered into the dendrogram, the branches that connect the genes describe a module. When the branch cutting is performed, the numbers of the modules will vary based on the branches, and hence will also the labels of the modules. This explains the differences between the same color modules in the networks.

The merging also made that some of the modules that are in one of the networks seem to not be present in the other network, but they are merged into another module. In Figure 24, the control blue module shows high preservation because this module was present in the case network before the merging, and is merged into the case black module, which also explains the high overlap in the correspondence matrix. This indicates that merging is also one of the reasons for the differences in the modules between the condition networks.

As discussed earlier the grey module also seemed to be interesting to study which gene is not assigned to a module in any of the condition networks, or identify the genes that were not in the

case grey module, but in the control grey module. The different modules of the networks might also have been interesting, and by only analyzing 2 modules out of 14 in the case and 21 in the control network, there is a risk of missing modules containing dysregulated pathways with causative genes for this disease.

5.4.2 ConsensusPathDB

The analysis with ConsensusPathDB gave an idea of what pathways the modules of the different conditions contained. A challenge in the analysis with this tool was the case black module, as it contained many genes compared to the other modules selected for analysis. This analysis and a match with the PD pathway verified that WGCNA methodology with the preservation statistics will help to identify dysregulated pathways.

During this study, a curiosity raised about if the PD related pathways and genes belonged to the case blue module or the case black module before merging. Both networks contain the same set of genes, which means that these PD genes must be in the control network too. The previously described similarity between the case black module and the control blue module indicates a high probability that the PD genes are in the control blue module and suggests that this module should have also been analyzed. At the same time, these two modules are the largest modules of the networks and contain approximately 50% of the genes. Analyzing modules that are this big could give an inaccurate indication by a variety of functionalities.

The pathways identified by ConsensusPathDB and the identification of the genes in the pathways resulted in some interesting genes, where some of them were found to be more studied than others. The genes that were more studied, and found to be associated with functionalities that could be related to PD are described in this study. The other genes that are mentioned, but not described, were not studied enough to relate to PD, and some barely had any information about the functionalities of the gene.

In this study, interesting pathways are pathways associated with characteristics of PD described in the background section 1.1, which suggests there might be other pathways than the ones analyzed in this study that are interesting to study in the case of PD.

5.4.3 Module Network Analysis

ClueGO was used to look at the interactions between the pathways identified in the case and control modules of pink and black. The pathways that were visualized here were also found with ConsensusPathDB, but with this visualization, it was easier to look for pathways with high similarity. This also revealed different functional groups, which helped in identifying central functions for the disease pathways. The case black module was also a challenge in this tool, but the interesting pathways found in ConsensusPathDB gave an idea of what pathways to look for. The risk of using the data from ConsensusPathDB to select out interesting pathways is that considering the level of details tuned in these networks. Maybe some pathways that were not found in ConsensusPathDB analysis could also be interesting to study. On the other side, there were fewer genes after filtering, which could give more specific pathway lists considering the genes with the highest connectivity.

Network topology is widely used to analyze biological interaction networks, especially for differential analysis, where betweenness centrality(BC) measures were used in this study. A challenge here was, by all the filtering of genes and their interactions, the black modules had no overlapping genes between the control and case modules that could be compared. The comparison of BC measures of the overlapping genes between the case and control pink modules revealed some genes that

were associated with dysregulated functions that are known in the PD brain. Many other network topology measures could have been used to identify interesting genes as well, in addition to BC measures.

5.4.4 The Interesting Genes

All the different analysis methods of the modules and their genes resulted in interesting genes. These genes are identified by many steps of filtering and analyses that resulted in 20 genes found to be interesting, from approximately 50 000 genes.

A study of changes in cell composition of control vs PD patients [60] revealed that there was a big difference in cell composition between parts of the brain that changes the most, both biological differences and technical variation in sample dissection and preparation. This study also suggested that the observed gene profiles within PD patients were influenced by differences in cellular composition and driven by technical factors associated with RNA quality [60]. This indicated that the gene profiles observed from PD patients in this study may be because of different cell composition in the bulk tissues, and because of technical variation in RNA quality. Nido *et al.* [60] also did a differential expression analysis that resulted in downplaying some characteristics associated with PD [60]. As suggested in this study with WGCNA as well, Nido *et al* [60] also found that the changes are in the number of neurons and not directly in the signal transmitting pathways. Down-regulation of mitochondrial pathways, for example, complex I, is found to be driven by altered cellular composition [60]. This analysis highlighted processes related to endoplasmatic reticulum(ER) and unfolded protein response, which is also mentioned in this study when describing *SEC63*. The challenge of identifying transcriptional events that happen because of changes in cellular composition by cell-type correction may be solved by single-cell or cell-sorting based methods [60].

5.5 Personalized Medicine

There is currently only one treatment for all PD patients, the dopamine replacement treatment(DRT). This treatment does not cure PD but decreases the motor symptoms by increasing levels of dopamine which strengthens the signal transmission for movement. As described PD is not only because of dopamine levels, many other factors play a role in this disease. "One treatment for all" will benefit for some patients, but it may also have a negative effect on some patients. A combination of symptoms, biological factors such as Lewy bodies and genetic risk factors are used to diagnose PD, and to differ this disease from for example Alzheimers. For a complex multi-system disease like PD, a "cocktail therapy" involving all the parts of P4 medicine is thought to be more sufficient than "one treatment for all". This study of using WGCNA and identifying causative genes in dysregulated pathways focus on genetic risk factors in the predictive part.

Many researchers are working to find genetic risk factors and early signs related to PD, but it is very difficult as this is a disease common by aging, a complicated process itself. Oxidative stress which is described as a biological factor for developing PD is also a factor of the aging process. It is also known that when the motor symptoms are observed it might be too late for treatment.

With time it is less time consuming to get an output of the genetic risk factors for a patient, but even though a patient has a probability for getting PD, they might have a probability for other diseases as well. Which of the diseases should be looked into and how many biological dysfunctionalities will be looked for every single patient? The list of genes related to PD is ever-increasing because of the multiple factors associated with PD. The genes identified as interesting genes in this study might be validated through other studies and might be useful for the study of

PD, also in personalized medicine in PD. Genetic risk factors can help in predicting the disease in an earlier stage of the disease, also by combining non-motor symptoms as early signs of PD.

There are different genes in every biological dysfunctionality that may lead to PD by mutations, and some dysfunctionalities are also a part of other similar diseases like Alzheimer's. Available data of PD patients, for example, earlier blood tests can help to study earlier changes on a molecular level, which then can be used for predicting PD before the symptoms take place.

Cerebrospinal fluid(CSF) has also been used in studying biomarkers that were found to differentiate PD patients from healthy persons and similar diseases like Alzheimer's disease [61]. The study of CSF biomarkers also found that it could be used to differentiate between subgroups of PD and to monitor PD progression in longitudinal samples. Subgroups can then be used in the predictive part to indicate these changes in the molecular functions of PD patients much earlier, or it can be used for developing treatments more personalized for these subgroups before considering each individual. The identified genes from this study can be studied as biomarkers in CSF.

5.6 Further Study

This methodology and the hypotheses suggested in the study of diabetes type 1 [42] have shown results of some dysregulated pathways and interesting genes to study further. The tuned parameters can be changed and result in more modules with fewer genes in each module, which might give more specific modules associated with biological functions. Many more network topology properties can be studied with the networks visualized in Cytoscape that may reveal other interesting genes too. More modules were different between the conditions that could be interesting to look into, for example, the control blue module and the grey modules. The genes identified in this study can be studied further to associate these genes as possible genetic risk factors of PD, and this methodology for identifying dysregulated pathways and causative genes can be applied to other disease data as well.

6 Conclusion

Parkinson's disease is a complex progressive multi-system disease common by age. It is a very difficult disease to prevent or predict, as there are multiple biological factors as well as environmental factors triggering and causing this disease. To study such complex diseases many resources of different research fields are required, where one could be the genetic analysis to find genetic risk factors that may help to predict the disease. This study with a weighted gene co-expression network(WGCNA) methodology is an approach to identify causative genes by looking at dysregulated pathways in modules defined by clusters within co-expression s. Many methodologies to do this kind of network analysis, but for differential network analysis, WGCNA methodology is one of the most widely used methodologies.

WGCNA contains multiple functions and methods for filtering the data, calculating the correlations between gene profiles, constructing the network, identifying the modules, and visualization with more. ConsensusPathDB was used to identify dysregulated pathways, and Cytoscape was used to study network topology properties and to identify pathways and their interactions with ClueGO plug-in. These analyses resulted in a set of genes related to known dysregulation in PD patients, and that dysfunctions of synapses and synaptic transmissions may not play an important role for PD patients.

One of the challenges in this study was to tune the parameters in automatic functions, which also make it challenging to reproduce the results. With the parameters set, this study resulted in many interesting pathways and genes that were associated with dysfunctions of PD. The networks resulted in differences in the modules that could be challenging when comparing the different networks. Lastly, eliminating modules and genes to select only a few for further analysis was also a challenge, as some modules or genes that could be interesting may be left out.

This study was done as a contribution to the study of personalized medicine in PD, especially focusing on finding genes that might be genetic risk factors within PD patients, by doing a differential expression analysis between healthy controls and PD patients. It is important to keep in mind is that as this is a complex multi-system disease, all parts of P4 medicine are equally important to consider when designing a treatment. Dopamine replacement treatment(DRT) may be beneficial for some patients, but as PD is different for each patient, the treatments should also be accordingly for greater chances for a positive effect for all patients. "One treatment for all" such as DRT is a good start towards personalized medicine in PD. This treatment can be modified by pharmacology and improve the effect of the treatment for more patients.

Glossary

BC	Betweenness centrality.
DRT	Dopamine replacement treatment.
EOPD	Early- onset Parkinson's Disease.
ER	endoplasmic reticulum.
GO	Gene Ontology.
GWAS	Genome- wide Association Studies.
hub genes	central genes highly connected with other genes.
KEGG	Kyoto Encyclopedia for genes and genomes.
LB	Lewy Bodies.
MAD	Median Absolute Deviation.
median absolute deviation	Measure of variability of univariate sample of quantitative data.
modules	the branches of a gene dendrogram, clusters of gene-expression profiles.
monogenic	Single gene causing disease.
MPTP	1-methyl-4-phenyl-1,2,3,6-tetrahydropyrodine.

mtDNA	mitochondrial DNA.
NBB	Cohort; Netherlands Brain Bank.
neurodegeneration	progressive neuronal loss in structure or function and neuronal death.
neuromelanin	A dark pigmentation composed of proteins, lipids, and products of the DA metabolism.
NGS	Next generation sequencing.
PA	Cohort; Poly- A capture RNA seq.
PD	Parkinson's Disease.
proteasome	Protein complexes which degrade damaged proteins.
PW	Cohort; the Norwegian Park West Study.
ROS	Reactive oxygen species.
scale-free property distribution	Where the distribution of node degrees follows a power law.
SNpc	Substantia Nigra parsa compacta.
TOM	Topological overlap matrix.
WGCNA	Weighted Gene Correlation Network analysis.

References

[1] W. Dauer and S. Przedborski, "Parkinson's disease: mechanisms and models," *neuron*, vol. 39, no. 6, pp. 889–909, 2003.

[2] G. Borrageiro, W. Haylett, S. Seedat, H. Kuivaniemi, and S. Bardien, "A review of genome-wide transcriptomics studies in parkinson's disease," *European Journal of Neuroscience*, vol. 47, no. 1, pp. 1–16, 2018.

[3] A. Reeve, E. Simcox, and D. Turnbull, "Ageing and parkinson's disease: why is advancing age the biggest risk factor?," *Ageing research reviews*, vol. 14, pp. 19–30, 2014.

[4] C. Klein and A. Westenberger, "Genetics of parkinson's disease," *Cold Spring Harbor perspectives in medicine*, vol. 2, no. 1, p. a008888, 2012.

[5] K. J. Barnham, C. L. Masters, and A. I. Bush, "Neurodegenerative diseases and oxidative stress," *Nature reviews Drug discovery*, vol. 3, no. 3, p. 205, 2004.

[6] P. M. McLendon and J. Robbins, "Proteotoxicity and cardiac dysfunction," *Circulation research*, vol. 116, no. 11, pp. 1863–1882, 2015.

[7] A. Singleton and J. Hardy, "Progress in the genetic analysis of parkinson's disease," *Human molecular genetics*, vol. 28, no. R2, pp. R215–R218, 2019.

[8] C. Blauwendraat, M. A. Nalls, and A. B. Singleton, "The genetic architecture of parkinson's disease," *The Lancet Neurology*, vol. 19, no. 2, pp. 170–178, 2020.

[9] R. Lowe, N. Shirley, M. Bleackley, S. Dolan, and T. Shafee, "Transcriptomics technologies," *PLoS computational biology*, vol. 13, no. 5, 2017.

[10] T. Kavanagh, J. D. Mills, W. S. Kim, G. M. Halliday, and M. Janitz, "Pathway analysis of the human brain transcriptome in disease," *Journal of Molecular Neuroscience*, vol. 51, no. 1, pp. 28–36, 2013.

[11] J. Kelly, R. Moyeed, C. Carroll, D. Albani, and X. Li, "Gene expression meta-analysis of parkinson's disease and its relationship with alzheimer's disease," *Molecular brain*, vol. 12, no. 1, p. 16, 2019.

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

[13] F. Saudou, S. Finkbeiner, D. Devys, and M. E. Greenberg, "Huntingtin acts in the nucleus to induce apoptosis but death does not correlate with the formation of intranuclear inclusions," *Cell*, vol. 95, no. 1, pp. 55–66, 1998.

[14] C. J. Cummings, E. Reinstein, Y. Sun, B. Antalffy, Y.-h. Jiang, A. Ciechanover, H. T. Orr, A. L. Beaudet, and H. Y. Zoghbi, "Mutation of the e6-ap ubiquitin ligase reduces nuclear inclusion frequency while accelerating polyglutamine-induced pathology in sca1 mice," *Neuron*, vol. 24, no. 4, pp. 879–892, 1999.

[15] W. D. Parker Jr, S. J. Boyson, and J. K. Parks, "Abnormalities of the electron transport chain in idiopathic parkinson's disease," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 26, no. 6, pp. 719–723, 1989.

[16] B. Halliwell and J. M. Gutteridge, *Free radicals in biology and medicine*. Oxford University Press, USA, 2015.

[17] G. Cohen, "Oxidative stress, mitochondrial respiration, and parkinson's disease," *Annals of the New York Academy of Sciences*, vol. 899, no. 1, pp. 112–120, 2000.

[18] B. Bilgic, A. Pfefferbaum, T. Rohlfing, E. V. Sullivan, and E. Adalsteinsson, "Mri estimates of brain iron concentration in normal aging using quantitative susceptibility mapping," *Neuroimage*, vol. 59, no. 3, pp. 2625–2635, 2012.

[19] A. Daugherty and N. Raz, "Age-related differences in iron content of subcortical nuclei observed in vivo: a meta-analysis," *Neuroimage*, vol. 70, pp. 113–121, 2013.

[20] D. Dexter, F. Wells, A. Lee, F. Agid, Y. Agid, P. Jenner, and C. Marsden, "Increased nigral iron content and alterations in other metal ions occurring in brain in parkinson's disease," *Journal of neurochemistry*, vol. 52, no. 6, pp. 1830–1836, 1989.

[21] E. M. Haacke, Y. Miao, M. Liu, C. A. Habib, Y. Katkuri, T. Liu, Z. Yang, Z. Lang, J. Hu, and J. Wu, "Correlation of putative iron content as represented by changes in r2* and phase with age in deep gray matter of healthy adults," *Journal of Magnetic Resonance Imaging*, vol. 32, no. 3, pp. 561–576, 2010.

[22] E. Sofic, W. Paulus, K. Jellinger, P. Riederer, and M. Youdim, "Selective increase of iron in substantia nigra zona compacta of parkinsonian brains," *Journal of neurochemistry*, vol. 56, no. 3, pp. 978–982, 1991.

[23] A. Friedman, J. Galazka-Friedman, and E. R. Bauminger, "Iron as a trigger of neurodegeneration in parkinson's disease," *Handbook of clinical neurology*, vol. 83, pp. 493–505, 2007.

[24] B. Picconi, G. Piccoli, and P. Calabresi, "Synaptic dysfunction in parkinson's disease," in *Synaptic Plasticity*, pp. 553–572, Springer, 2012.

[25] T. Schirinzi, G. Madeo, G. Martella, M. Maltese, B. Picconi, P. Calabresi, and A. Pisani, "Early synaptic dysfunction in parkinson's disease: insights from animal models," *Movement Disorders*, vol. 31, no. 6, pp. 802–813, 2016.

[26] R.-S. Wang, B. A. Maron, and J. Loscalzo, "Systems medicine: evolution of systems biology from bench to bedside," *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, vol. 7, no. 4, pp. 141–161, 2015.

[27] S. van Dam, U. Vosa, A. van der Graaf, L. Franke, and J. P. de Magalhaes, "Gene co-expression analysis for functional classification and gene–disease predictions," *Briefings in bioinformatics*, vol. 19, no. 4, pp. 575–592, 2017.

[28] P. Langfelder and S. Horvath, "Wgcna: an r package for weighted correlation network analysis," *BMC bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[29] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis," *Statistical applications in genetics and molecular biology*, vol. 4, no. 1, 2005.

[30] S. Ballouz, W. Verleyen, and J. Gillis, "Guidance for rna-seq co-expression network construction and analysis: safety in numbers," *Bioinformatics*, vol. 31, no. 13, pp. 2123–2130, 2015.

[31] J. P. Doye, "Network topology of a potential energy landscape: A static scale-free network," *Physical review letters*, vol. 88, no. 23, p. 238701, 2002.

[32] A. M. Yip and S. Horvath, "Gene network interconnectedness and the generalized topological overlap measure," *BMC bioinformatics*, vol. 8, no. 1, p. 22, 2007.

[33] G. O. Consortium, "The gene ontology (go) database and informatics resource," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D258–D261, 2004.

[34] P. Langfelder and S. Horvath, "Fast R functions for robust correlations and hierarchical clustering," *Journal of Statistical Software*, vol. 46, no. 11, pp. 1–17, 2012.

[35] E. A. Serin, H. Nijveen, H. W. Hilhorst, and W. Ligterink, "Learning from co-expression networks: possibilities and challenges," *Frontiers in plant science*, vol. 7, p. 444, 2016.

[36] G. O. Consortium, "The gene ontology resource: 20 years and still going strong," *Nucleic acids research*, vol. 47, no. D1, pp. D330–D338, 2019.

[37] G. O. Consortium, "Gene ontology consortium: going forward," *Nucleic acids research*, vol. 43, no. D1, pp. D1049–D1056, 2015.

[38] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[39] A. Kamburov, U. Stelzl, H. Lehrach, and R. Herwig, "The consensuspathdb interaction database: 2013 update," *Nucleic acids research*, vol. 41, no. D1, pp. D793–D800, 2013.

[40] M. Kanehisa, Y. Sato, M. Furumichi, K. Morishima, and M. Tanabe, "New approach for understanding genome variations in kegg," *Nucleic acids research*, vol. 47, no. D1, pp. D590–D595, 2019.

[41] J. Ruan, A. K. Dean, and W. Zhang, "A general co-expression network-based approach to gene expression analysis: comparison and applications," *BMC systems biology*, vol. 4, no. 1, p. 8, 2010.

[42] I. R. Medina and Z. Lubovac-Pilav, "Gene co-expression network analysis for identifying modules and functionally enriched pathways in type 1 diabetes," *PloS one*, vol. 11, no. 6, 2016.

[43] P. Langfelder, R. Luo, M. C. Oldham, and S. Horvath, "Is my network module preserved and reproducible?," *PLoS computational biology*, vol. 7, no. 1, 2011.

[44] S. L. Carter, C. M. Brechbühler, M. Griffin, and A. T. Bond, "Gene co-expression network topology provides a framework for molecular characterization of cellular state," *Bioinformatics*, vol. 20, no. 14, pp. 2242–2250, 2004.

[45] D. Koschützki and F. Schreiber, "Centrality analysis methods for biological networks and their application to gene regulatory networks," *Gene regulation and systems biology*, vol. 2, pp. GRSB–S702, 2008.

[46] M. Ray and W. Zhang, "Analysis of alzheimer's disease severity across brain regions by topological analysis of gene co-expression networks," *BMC systems biology*, vol. 4, no. 1, p. 136, 2010.

[47] N. Titova and K. R. Chaudhuri, "Personalized medicine in parkinson's disease: time to be precise," *Movement Disorders*, vol. 32, no. 8, p. 1147, 2017.

[48] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0.

[49] E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock, "Go:: Termfinder—open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes," *Bioinformatics*, vol. 20, no. 18, pp. 3710–3715, 2004.

[50] F. Supek, M. Bošnjak, N. Škunca, and T. Šmuc, "Revigo summarizes and visualizes long lists of gene ontology terms," *PloS one*, vol. 6, no. 7, 2011.

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

[52] J. Dong and S. Horvath, "Understanding network concepts in modules," *BMC systems biology*, vol. 1, no. 1, p. 24, 2007.

[53] P. Langfelder and S. Horvath, "Eigengene networks for studying the relationships between co-expression modules," *BMC systems biology*, vol. 1, no. 1, p. 54, 2007.

[54] A. Kamburov, C. Wierling, H. Lehrach, and R. Herwig, "Consensuspathdb—a database for integrating human functional interaction networks," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D623–D628, 2009.

[55] A. Kamburov, K. Pentchev, H. Galicka, C. Wierling, H. Lehrach, and R. Herwig, "Consensuspathdb: toward a more complete picture of cell biology," *Nucleic acids research*, vol. 39, no. suppl_1, pp. D712–D717, 2011.

[56] G. Bindea and B. Mlecnik, "Laboratory of integrative cancer immunology umrs1138 cordeliers research center, paris, france,"

[57] G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. I. Stein, R. Nudel, I. Lieder, Y. Mazor, *et al.*, "The genecards suite: from gene data mining to disease genome sequence analyses," *Current protocols in bioinformatics*, vol. 54, no. 1, pp. 1–30, 2016.

[58] J. Smeitink, L. van den Heuvel, and S. DiMauro, "The genetics and pathology of oxidative phosphorylation," *Nature Reviews Genetics*, vol. 2, no. 5, pp. 342–352, 2001.

[59] C. Perier and M. Vila, "Mitochondrial biology and parkinson's disease," *Cold Spring Harbor perspectives in medicine*, vol. 2, no. 2, p. a009332, 2012.

[60] G. S. Nido, F. Dick, L. Toker, K. Petersen, G. Alves, O.-B. Tysnes, I. Jonassen, K. Haugarvoll, and C. Tzoulis, "Common gene expression signatures in parkinson's disease are driven by changes in cell composition," *BioRxiv*, p. 778910, 2019.

[61] M. Shi, J. Bradner, A. M. Hancock, K. A. Chung, J. F. Quinn, E. R. Peskind, D. Galasko, J. Jankovic, C. P. Zabetian, H. M. Kim, *et al.*, "Cerebrospinal fluid biomarkers for parkinson disease diagnosis and progression," *Annals of neurology*, vol. 69, no. 3, pp. 570–580, 2011.

A Appendix

A.1 ConsensusPathDB outputs

- Black-case
- Black-control
- Pink-case
- Pink-control
- dark grey-control
- dark orange-control
- Orange-Control

p-value	q-value	pathway	source	external_id	*members_input_overlap*
3,10E-25	9,50E-23	Oxidative phosphorylation - Homo sapiens (human)	KEGG	path:hsa00190	*MT-CO1; MT-CO2; MT-CO3; NDUFB9; NDUFB8; NDUFB7; NDUFB6; NDUFB5; NDUFB4; NDUFB3; NDUFB2; NDUFB1; UQCRQ; UQCRH; UQCRB; NDUFAB1; MT-ND1; MT-ND2; MT-ND3; MT-ND4; MT-ND5; MT-ND6; ATP6AP1; NDUFB11; NDUFB10; NDUFA10; NDUFA11; NDUFA12; NDUFA13; MTATP6; COX6A1; UQCRFS1; UQCRC2; UQCRC1; ATP6V1D; ATP6V1F; ATP6V1A; ATP6V1H; COX6B1; MT-ND4L; COX6C; MT-CYB; COX8A; COX7A1; COX7A2; ATP6V0A1; ATP6V1G1; COX11; COX15; COX17; PPA2; PPA1; NDUFV3; NDUFV2; NDUFV1; MT-ATP8; NDUFA6; ATP6V0C; NDUFA4; NDUFA5; NDUFA2; NDUFA3; NDUFA1; NDUFA8; NDUFA9; COX4I1; ATP6V0D1; ATP6V1E1; COX7B; COX7C; ATP6V0B; COX5A; COX5B; UQCR10; UQCR11; SDHC; ATP6V1B2; ATP6V1G2; NDUFC1; NDUFC2; ATP6V0A2; NDUFS8; ATP6V1C1; COX7A2L; CYC1; NDUFS1; NDUFS2; NDUFS3; NDUFS4; NDUFS5; NDUFS6; NDUFS7; SDHA; ATP6V0E1; ATP6V0E2; SDHB; SDHD*
2,81E-23	4,30E-21	Parkinson disease - Homo sapiens (human)	KEGG	path:hsa05012	*MT-CO1; MT-CO2; MT-CO3; UBE2J1; NDUFB9; NDUFB8; NDUFB7; NDUFB6; NDUFB5; NDUFB4; NDUFB3; NDUFB2; NDUFB1; UQCRQ; UQCRH; UQCRB; NDUFAB1; MT-ND1; MT-ND2; MT-ND3; MT-ND4; MT-ND5; MT-ND6; PARK7; NDUFB11; NDUFB10; MT-ATP8; NDUFA11; NDUFA12; NDUFA13; MT-ATP6; SLC25A6; SLC25A5; SLC25A4; COX6A1; ADCY5; UBB; UQCRFS1; UQCRC2; UQCRC1; UBE2L3; SEPT5; COX6B1; ADORA2A; LRRK2; MT-ND4L; COX6C; MT-CYB; COX8A; UBA1; COX7A1; COX7A2; SNCA; GNAL; VDAC3; VDAC2; VDAC1; NDUFV3; NDUFV2; NDUFV1; NDUFA10; NDUFA6; NDUFA4; NDUFA5; NDUFA2; NDUFA3; NDUFA1; NDUFA8; NDUFA9; COX4I1; UCHL1; PRKACA; PRKACB; PPIF; COX7B; COX7C; COX5A; COX5B; UQCR10; UQCR11; CYCS; PINK1; UBE2G1; NDUFC1; NDUFC2; NDUFS8; COX7A2L; CYC1; NDUFS1; NDUFS2; NDUFS3; NDUFS4; NDUFS5; NDUFS6; NDUFS7; SDHA; SDHC; SDHB; SDHD*
6,82E-22	6,96E-20	Thermogenesis - Homo sapiens (human)	KEGG	path:hsa04714	*SMARCC2; MT-CO1; MT-CO2; MT-CO3; ACTG1; NDUFB9; NDUFB8; NDUFB7; NDUFB6; NDUFB5; NDUFB4; NDUFB3; NDUFB2; NDUFB1; MT-ND4L; SMARCC1; UQCRH; UQCRB; NDUFAB1; MT-ND1; MT-ND2; MT-ND3; MT-ND4; MT-ND5; MT-ND6; NDUFB11; NDUFB10; NDUFA10; NDUFA11; NDUFA12; NDUFA13; MT-ATP6; FGFR1; PRKAG2; CREB1; SMARCD3; SMARCD1; UQCRQ; COX6A1; KRAS; ADCY5; ADCY1; ADCY3; ACSL3; ACSL1; ACSL6; ACSL4; GRB2; ACTB; UQCRC2; UQCRC1; NDUFAF3; ADCY8; HRAS; NDUFC2; COX6B1; PPARGC1A; PRKG1; ZNF516; COX6C; MT-CYB; COX8A; COX7A1; COX7A2; UQCRFS1; RPS6KA2; RPS6KA3; GNAS; COX18; COX19; COX11; KDM3A; COX14; COX15; COX16; COX17; DPF3; NDUFAF5; NDUFAF4; NDUFAF1; MAPK12; KDM1A; TSC2; TSC1; NDUFV3; NDUFV2; NDUFV1; RPS6KB1; MT-ATP8; NDUFA6; NDUFA4; NDUFA5; NDUFA2; NDUFA3; NDUFA1; NDUFA8; NDUFA9; COX4I1; MTOR; PRKACA; PRKACB; RHEB; COX7B; COX7C; COX5A; COX5B; UQCR10; UQCR11; PRKAA2; PRKAA1; SMARCA2; SMARCA4; RPTOR; NRAS; NDUFC1; CPT1C; CPT1B; SMARCB1; NDUFS6; COA5; COA7; COA6; COA3; SDHA; COX7A2L; CYC1; NDUFS1; NDUFS2; NDUFS3; NDUFS4; NDUFS5; ACTL6B; NDUFS7; NDUFS8; SDHC; SDHB; SDHD*

2,16E-18	1,65E-16	Huntington disease - Homo sapiens (human)	KEGG	path:hsa05016	*MT-CO1; TGM2; MT-CO3; NDUFB9; NDUFB8; NDUFB7; NDUFB6; NDUFB5; NDUFB4; NDUFB3; NDUFB2; NDUFB1; UQCRQ; UQCRH; BDNF; UQCRB; NDUFAB1; NDUFB11; NDUFB10; NDUFA10; NDUFA11; NDUFA12; NDUFA13; MT-ATP6; SOD2; IFT57; SOD1; SP1; SLC25A6; SLC25A5; SLC25A4; COX6A1; UQCRFS1; UQCRC2; UQCRC1; VDAC3; MT-CO2; CLTA; CLTC; CLTB; COX6B1; AP2B1; AP2S1;*
					PPARGC1A; POLR2J2; COX6C; PLCB1; PLCB4; MT-CYB; COX8A; POLR2E; POLR2A; POLR2C; POLR2B; COX7A1; COX7A2; ITPR1; POLR2I; POLR2K; TBPL1; GNAQ; GPX1; CREBBP; POLR2J3; HDAC2; DCTN2; VDAC2; VDAC1; DCTN1; DCTN4; NDUFV3; NDUFV2; NDUFV1; AP2M1; MT-ATP8; NDUFA6; NDUFA4; NDUFA5; NDUFA2; NDUFA3; NDUFA1; NDUFA8; NDUFA9; COX4I1; PPIF; COX7B; COX7C; COX5A; COX5B; UQCR10; UQCR11; CYCS; NDUFC1; NDUFC2; TFAM; NDUFS8; AP2A1; CREB1; AP2A2; SDHC; COX7A2L; CYC1; GRIN2B; NDUFS1; NDUFS2; NDUFS3; NDUFS4; NDUFS5; NDUFS6; NDUFS7; SDHA; DNAL4; SDHB; SDHD; DNAL1
1,19E-17	7,28E-16	Alzheimer disease - Homo sapiens (human)	KEGG	path:hsa05010	*MT-CO1; MT-CO2; MT-CO3; NDUFB9; NDUFB8; NDUFB7; NDUFB6; NDUFB5; NDUFB4; NDUFB3; NDUFB2; NDUFB1; UQCRQ; UQCRH; UQCRB; NDUFAB1; CALM2; CALM3; CALM1; APBB1; NDUFB11; NDUFB10; NDUFA10; CAPN2; CAPN1; NDUFA13; MT-ATP6; LPL; COX6A1; MT-ATP8; APP; GSK3B; UQCRFS1; UQCRC2; UQCRC1; ATP2A2; NOS1; GAPDH; COX6B1; COX6C; PLCB1; PLCB4; MT-CYB; COX8A; COX7A1; COX7A2; ITPR1; GNAQ; TNFRSF1A; SNCA; PPP3R1; ITPR3; NDUFV3; NDUFV2; NDUFV1; MAPK1; NAE1; NDUFA11; NDUFA6; NDUFA4; NDUFA5; NDUFA2; NDUFA12; NDUFA1; NDUFA8; NDUFA9; COX4I1; EIF2AK3; CDK5; COX7B; COX7C; COX5A; COX5B; UQCR10; UQCR11; CYCS; BAD; NDUFA3; GRIN2C; BID; NDUFS3; NDUFC1; NDUFC2; RTN4; RTN3; NDUFS8; SDHC; COX7A2L; BACE2; SDHB; CYC1; GRIN2B; NDUFS1; NDUFS2; GRIN2A; NDUFS4; NDUFS5; NDUFS6; NDUFS7; SDHA; PPP3CA; PPP3CB; ATF6; SDHD*
2,28E-16	1,16E-14	Non-alcoholic fatty liver disease (NAFLD) - Homo sapiens (human)	KEGG	path:hsa04932	*MT-CO1; MT-CO2; MT-CO3; NDUFB9; NDUFB8; NDUFB7; NDUFB6; NDUFB5; NDUFB4; NDUFB3; NDUFB2; NDUFB1; UQCRQ; IKBKB; AKT3; UQCRH; UQCRB; NDUFAB1; PIK3CA; PIK3CB; NDUFB11; NDUFB10; NDUFA10; NDUFA11; NDUFA12; NDUFA13; RELA; PRKAG2; COX6A1; GSK3A; GSK3B; UQCRFS1; UQCRC2; UQCRC1; COX6B1; PIK3R3; PIK3R1; COX6C; MT-CYB; COX8A; COX7A1; COX7A2; TNFRSF1A; MAPK10; NDUFV3; NDUFV2; NDUFV1; ADIPOR1; MAPK8; MAPK9; NDUFA6; NDUFA4; NDUFA5; NDUFA2; NDUFA3; NDUFA1; RXRA; NDUFA8; NDUFA9; COX4I1; MLXIP; CYCS; EIF2AK3; COX7B; COX7C; CDC42; COX5A; COX5B; UQCR10; UQCR11; PRKAA2; PRKAA1; EIF2S1; BID; NDUFC1; NDUFC2; RAC1; NDUFS8; IL6R; COX7A2L; CYC1; NDUFS1; NDUFS2; NDUFS3; NDUFS4; NDUFS5; NDUFS6; NDUFS7; SDHA; SDHC; SDHB; SDHD*

1,88E-15	8,20E-14	Retrograde endocannabinoid signaling - Homo sapiens (human)	KEGG	path:hsa04723	*NDUFB9; NDUFB8; NDUFB7; NDUFB6; NDUFB5; NDUFB4; NDUFB3; NDUFB2; NDUFB1; NDUFAB1; MT-ND1; MT-ND2; MT-ND3; MT-ND4; MT-ND5; MT-ND6; GNG10; GNG11; NDUFB11; NDUFB10; NDUFA10; NDUFA11; NDUFA12; NDUFA13; ABHD6; ADCY5; ADCY1; ADCY3; ADCY8; GNG4; GNG2; GNG3; SLC17A6; GRIA2; GABRG2; GRIA4; MT-ND4L; PLCB1; PLCB4; ITPR3; ITPR1; GNAQ; GABRD; SLC32A1; NAPEPLD; MAPK10; MAPK12; GABRA5; GABRA4; GABRA1; GABRA3; NDUFV3; NDUFV2; NDUFV1; MAPK1; GRIA3; GNAO1; MAPK8; MAPK9; RIMS1; DAGLA; NDUFA4; NDUFA5; NDUFA2; NDUFA3; NDUFA1; GNB5; GNB1; NDUFA8; NDUFA9; GABRB2; KCNJ3; FAAH; KCNJ9; SLC17A7; PRKACA; PRKACB; NDUFA6; NDUFC1; NDUFC2; PRKCB; PRKCG; NDUFS1; NDUFS2; NDUFS3; NDUFS4; NDUFS5; NDUFS6; NDUFS7; NDUFS8*
3,98E-14	1,52E-12	Protein processing in endoplasmic reticulum - Homo sapiens (human)	KEGG	path:hsa04141	*RPN1; RPN2; UBE2J1; TUSC3; HSPA8; MAN1B1; MARCH6; CANX; HYOU1; STUB1; HSP90AB1; NSFL1C; CAPN1; CUL1; DNAJC10; ERLEC1; SEC24A; EDEM3; SEC24C; SEC24B; MAN1A1; DNAJA2; UBE4B; DNAJA1; MBTPS2; MBTPS1; RNF185; RBX1; RNF5; EDEM1; UBE2D2; UBE2D3; HSP90B1; UBE2D1; NGLY1; DERL1; DNAJB11; UBXN6; OS9; ERP29; SEC13; HSPA4L; STT3B; STT3A; DNAJC5; DNAJC3; ATXN3; GANAB; BAG1; RAD23B; RAD23A; FBXO2; MAPK10; NPLOC4; SEC23A; SEC23B; SKP1; MAPK8; MAPK9; UGGT2; UGGT1; SEL1L; AMFR; SEC62; SAR1A; SEC63; BCAP31; EIF2AK1; EIF2AK3; EIF2AK2; EIF2AK4; PREB; CRYAB; HSPA1A; DDOST; SVIP; EIF2S1; SEC31B; SEC31A; UBQLN2; UBQLN1; UBQLN4; HSPBP1; UBE2G1; PDIA6; PDIA4; PDIA3; VCP; PLAA; CALR; CAPN2; SSR4; CKAP4; ATF6; SEC61A2*
1,47E-12	5,00E-11	Ubiquitin mediated proteolysis - Homo sapiens (human)	KEGG	path:hsa04120	*UBE2Q1; UBE3A; UBE2J1; WWP1; UBA6; BIRC6; HUWE1; UBE2D2; UBE2D3; RNF7; UBE2D1; FBXO2; UBE2Q2; UBE2L3; RCHY1; TRIM37; SMURF2; CDC16; FBXW11; RHOBTB1; SKP1; STUB1; UBE3C; MAP3K1; ANAPC11; CDC27; ANAPC13; CUL4B; UBR5; UBE2G1; TRIM32; PML; CUL5; DDB1; CUL1; CUL2; CUL3; SMURF1; UBA2; SIAH1; UBE3B; UBE2H; FANCL; UBE2E1; UBE4A; UBE2R2; UBE2E2; HERC1; HERC3; HERC2; RHOBTB2; UBE2E3; UBA1; ANAPC1; PIAS1; TRIP12; ANAPC5; ANAPC4; ANAPC7; BTRC; UBE2I; UBE4B; UBE2K; CUL7; UBE2M; UBE2N; UBE2A; FZR1; XIAP; UBE2F; CUL4A; PRPF19; SAE1; UBE2Z; UBE2B; BIRC2; RBX1; FBXW7; KLHL9; UBA3*
3,06E-12	9,36E-11	Proteasome - Homo sapiens (human)	KEGG	path:hsa03050	*PSMD8; PSMD4; PSMF1; PSMD6; PSMD1; PSMD3; PSMD2; PSMD7; PSMA2; PSMA3; PSMA1; PSMA6; PSMA7; PSMA4; PSMA5; PSMC1; PSMC2; PSMC3; ADRM1; PSMC5; PSMC6; POMP; PSME4; PSME3; PSMC4; PSMD11; PSMD13; PSMD12; PSMD14; PSMB7; PSMB6; PSMB5; PSMB4; PSMB3; PSMB2; PSMB1*

5,22E-11	1,45E-09	Endocytosis - Homo sapiens (human)	KEGG	path:hsa04144	*VPS4B; WWP1; HSPA8; CHMP2B; IST1; CHMP2A; CAV2; SMAP1; SMAP2; RAB4A; PARD6B; PARD6A; VPS35; IGF1R; RAB5A; RAB5B; SH3GLB2; VPS26A; ZFYVE9; KIF5A; ARPC1A; RAB11FIP4; RAB11FIP2; UBB; ARPC1B; GBF1; SNX3; CAPZB; KIF5C; SNX4; HRAS; CLTA; CLTC; CLTB; IQSEC1; IQSEC3; STAM; AP2B1; SMURF1; AP2S1; SMURF2; VPS37A; RAB11A; RAB11B; CHMP5; VTA1; RAB7A; ARPC3; ARPC2; RNF41; ARPC4; ARPC5; EPS15L1; AMPH; DNAJC6; TSG101; PDGFRA; EHD3; EHD2; SNX2; VPS26B; ARRB2; ARRB1; CHMP3; CHMP7; SH3GL3; SH3GL2; VPS45; SNX12; ARF3; ARF1; ARF5; AP2M1; RAB22A; RABEP1; PML; RAB10; VPS4A; ARFGEF2; CHMP1B; CHMP1A; CAPZA2; CAPZA1; ZFYVE27; RAB11FIP5; VPS36; PIP5K1C; PIP5K1B; ARPC5L; DNM1; ASAP2; CDC42; USP8; DNM3; CHMP4A; CHMP4B; ARFGAP1; HSPA1A; WASL; EPS15; RAB35; ARAP1; EEA1; STAM2; VPS29; SH3KBP1; PRKCI; HGS; ACAP3; AP2A1; SNF8; AP2A2; CYTH3; CYTH2; PARD3; PDCD6IP; ARFGAP2; WIPF2; VPS28; ARFGEF1*
6,54E-11	1,67E-09	Autophagy - animal - Homo sapiens (human)	KEGG	path:hsa04140	*BCL2L1; VMP1; IGF1R; MAP2K1; UVRAG; HMGB1; STK11; IGBP1; PRKAA1; MAP2K2; BAD; AKT3; CTSB; MAPK10; ZFYVE1; RB1CC1; TSC2; TSC1; RPTOR; PIK3C3; GABARAPL1; CAMKK2; PIK3CA; PIK3CB; NRAS; RPS6KB1; MAP3K7; GABARAP; CFLAR; PIK3R4; PIK3R3; ATG7; PIK3R1; ATG5; ATG9A; MAPK8; NRBF2; DAPK3; HRAS; MAPK1; ULK2; RRAGD; RAB7A; DDIT4; RRAGC; RRAGB; WIPI2; AMBRA1; PRKACB; CTSD; MRAS; PRKAA2; EIF2S1; ATG12; ATG13; ATG14; KRAS; ITPR1; BNIP3; ATG2B; MTMR4; BECN1; PPP2CA; MAPK9; EIF2AK3; GABARAPL2; EIF2AK4; PTEN; RRAGA; PRKACA; ATG16L1; MTOR; RHEB*
1,22E-10	2,87E-09	Spliceosome - Homo sapiens (human)	KEGG	path:hsa03040	*PLRG1; PQBP1; HSPA8; TCERG1; SLU7; SF3B3; SRSF8; DHX38; PRPF8; HSPA1A; ZMAT2; DHX15; TXNL4A; PRPF3; RBMX; BCAS2; SART1; PRPF6; SRSF5; SNRNP200; SRSF7; SRSF1; SRSF3; AQR; PRPF19; PRPF18; PRPF31; U2AF2; SRSF9; HNRNPU; PRPF4; EFTUD2; LSM4; HNRNPK; RBM17; SF3A1; SF3A3; HNRNPC; SRSF2; TRA2B; HNRNPM; DHX8; CDC5L; SNRNP27; SMNDC1; HNRNPA3; SF3B1; PCBP1; U2SURP; WBP11; U2AF1L4; SNRPB2; THOC3; SNRPD3; THOC1; SF3B2; CWC15; SNRPD2; DDX23; RBMXL1; SNRNP40; SNW1; CDC40; SNRPA1; SNRPD1; PPIL1; PRPF38A; DDX42; PUF60; DDX46; DDX5; RBM25; RBM22; XAB2; EIF4A3*
1,97E-10	4,29E-09	Mitophagy - animal - Homo sapiens (human)	KEGG	path:hsa04137	*USP8; BCL2L1; HRAS; USP15; FOXO3; PGAM5; MAPK10; RHOT1; RHOT2; PINK1; BECN1; OPTN; GABARAPL1; TBK1; GABARAPL2; CALCOCO2; NRAS; GABARAP; TFE3; ATG5; CITED2; RELA; KRAS; RAB7A; FUNDC1; BNIP3L; SP1; AMBRA1; MAPK8; MRAS; ATG9A; CSNK2A2; CSNK2A1; BNIP3; TAX1BP1; MAPK9; BCL2L13; EIF2AK3; MFN2; MFN1; FIS1; UBB; CSNK2B; TBC1D15*
2,10E-10	4,29E-09	Synaptic vesicle cycle - Homo sapiens (human)	KEGG	path:hsa04721	*STX1A; ATP6V1D; ATP6V1F; ATP6V1A; DNM3; SLC32A1; UNC13A; ATP6V1H; STXBP1; STX1B; CLTA; CLTC; CLTB; AP2B1; ATP6V1B2; AP2S1; NSF; AP2M1; NAPA; ATP6V0D1; RIMS1; ATP6V0A1; ATP6V0B; ATP6V0C; ATP6V1G1; CPLX2; DNM1; CPLX1; AP2A1; UNC13C; RAB3A; AP2A2; ATP6V1C1; SNAP25; VAMP2; SYT1; SLC17A6; SLC17A7; ATP6V1E1; ATP6V1G2; ATP6V0A2; ATP6V0E1; ATP6V0E2*
1,21E-07	2,31E-06	Citrate cycle (TCA cycle) - Homo sapiens (human)	KEGG	path:hsa00020	*CS; MDH2; MDH1; FH; IDH3A; OGDH; IDH3B; IDH3G; PDHB; OGDHL; PDHA1; DLD; SUCLG1; ACO1; ACO2; DLST; SUCLA2; DLAT; SDHA; ACLY; SDHC; SDHB; SDHD*

4,38E-07	7,88E-06	Phosphatidylinositol signaling system - Homo sapiens (human)	KEGG	path:hsa04070	*OCRL; PI4KB; PI4KA; PLCG1; IMPAD1; PI4K2A; CDIPT; DGKZ; ITPR1; CALM2; CALM3; INPPL1; CALM1; PIK3CA; PIK3CB; DGKQ; INPP4A; ITPKA; DGKH; SYNJ1; SACM1L; PIK3R1; DGKB; PIK3R3; MTMR4; MTMR2; DGKD; PIP4K2B; PLCB1; PLCB4; PIKFYVE; PRKCB; PRKCG; IMPA1; MTMR6; ITPR3; PIK3C3; IP6K1; INPP5A; INPP5B; INPP5F; MTMR7; PIP5K1C; CDS1; CDS2; INPP5K; PPIP5K2; PTEN; MTMR1; PIP5K1B; PIP4K2C; DGKE; INPP5J*
2,76E-06	4,70E-05	Insulin signaling pathway - Homo sapiens (human)	KEGG	path:hsa04910	*CRKL; MKNK2; NRAS; HRAS; PRKAG2; PRKAA2; PRKAA1; IKBKB; MAP2K2; BAD; AKT3; MAPK10; RAPGEF1; SORBS1; CRK; PDE3B; TSC2; TSC1; RPTOR; PPP1R3F; PPP1R3E; CALM2; CALM3; INPPL1; CALM1; PIK3CA; PIK3CB; RPS6KB1; PPARGC1A; PHKB; PYGB; MAPK1; FASN; PIK3R1; PIK3R3; MAPK8; MAPK9; FLOT1; PRKCI; ACACA; SHC1; SHC2; RHOQ; HK1; PHKA2; MTOR; PRKAR1A; PRKAR1B; EIF4E; KRAS; PRKAR2A; EIF4E2; PPP1CB; EXOC7; PPP1CA; INPP5K; MAP2K1; GSK3B; PTPRF; PRKAR2B; GRB2; PRKACA; BRAF; FLOT2; PRKACB; RHEB*
5,81E-06	9,35E-05	mTOR signaling pathway - Homo sapiens (human)	KEGG	path:hsa04150	*SLC7A5; ATP6V1D; ATP6V1F; ATP6V1A; SEH1L; NRAS; HRAS; CAB39; ATP6V1H; PRKAA1; IKBKB; MAP2K2; MAP2K1; AKT3; MIOS; RRAGA; RICTOR; TSC2; NPRL3; TSC1; RPTOR; FZD4; LPIN1; FZD8; PIK3CA; PIK3CB; RPS6KB1; DVL1; PIK3R3; PIK3R1; RNF152; ATP6V1B2; EIF4B; MAPK1; BRAF; ULK2; RRAGD; IGF1R; STK11; DDIT4; RRAGC; RRAGB; SEC13; PRKCB; PRKCG; PRKAA2; EIF4E2; MTOR; EIF4E; WNT5A; KRAS; CAB39L; RPS6KA2; RPS6KA3; TTI1; LRP5; TNFRSF1A; ATP6V1E1; GSK3B; PTEN; GRB2; ATP6V1G2; ATP6V1G1; DEPDC5; LAMTOR4; LAMTOR5; ATP6V1C1; RHEB; STRADA; LAMTOR3*

p-value	q-value	pathway	source	external_id	members_input_overlap
0,000267833	0,030422763	Thyroid hormone signaling pathway - Homo sapiens (human)	KEGG	path:hsa04919	*CREBBP; MED13L; TSC2; PRKCG; PIK3R2; EP300; RAF1; NCOR1*
0,000463516	0,030422763	Renal cell carcinoma - Homo sapiens (human)	KEGG	path:hsa05211	*VHL; PIK3R2; EP300; RAF1; CREBBP; RAPGEF1*
0,000629436	0,030422763	Melanogenesis - Homo sapiens (human)	KEGG	path:hsa04916	*ADCY5; ADCY6; PRKCG; DVL3; EP300; RAF1; CREBBP*
0,001288041	0,046691489	Glutamatergic synapse - Homo sapiens (human)	KEGG	path:hsa04724	*ADCY5; ADCY6; GRIK2; PRKCG; HOMER2; SHANK2; GRM2*
0,002604814	0,051479343	Aldosterone synthesis and secretion - Homo sapiens (human)	KEGG	path:hsa04925	*DAGLA; ADCY5; ADCY6; CACNA1I; PRKCG; CACNA1G*
0,00274784	0,051479343	Long-term potentiation - Homo sapiens (human)	KEGG	path:hsa04720	*PPP1R1A; EP300; RAF1; PRKCG; CREBBP*
0,002849692	0,051479343	FoxO signaling pathway - Homo sapiens (human)	KEGG	path:hsa04068	*HOMER2; SMAD3; AGAP2; PIK3R2; EP300; RAF1; CREBBP*
0,002979423	0,051479343	Hepatocellular carcinoma - Homo sapiens (human)	KEGG	path:hsa05225	*PHF10; ARID2; ARID1A; PRKCG; SMAD3; DVL3; PIK3R2; RAF1*
0,00319527	0,051479343	HIF-1 signaling pathway - Homo sapiens (human)	KEGG	path:hsa04066	*MKNK2; PRKCG; VHL; PIK3R2; EP300; CREBBP*
0,003758153	0,054493216	Adherens junction - Homo sapiens (human)	KEGG	path:hsa04520	*BAIAP2; EP300; FER; CREBBP; SMAD3*
0,004934186	0,060252476	Notch signaling pathway - Homo sapiens (human)	KEGG	path:hsa04330	*DVL3; EP300; CREBBP; NUMBL*
0,004986412	0,060252476	Phospholipase D signaling pathway - Homo sapiens (human)	KEGG	path:hsa04072	*ADCY5; ADCY6; RAF1; PIK3R2; CYTH3; TSC2; GRM2*
0,005569148	0,06211742	Cholinergic synapse - Homo sapiens (human)	KEGG	path:hsa04725	*ADCY5; ADCY6; CACNA1B; KCNQ2; PRKCG; PIK3R2*
0,006887561	0,07133545	Cushing syndrome - Homo sapiens (human)	KEGG	path:hsa04934	*ADCY5; ADCY6; CACNA1I; KMT2D; KMT2A; CACNA1G; DVL3*
0,009173279	0,088675033	Dilated cardiomyopathy (DCM) - Homo sapiens (human)	KEGG	path:hsa05414	*CACNB3; ADCY5; ADCY6; ITGA3; SGCD*
0,010403564	0,0942823	MAPK signaling pathway - Homo sapiens (human)	KEGG	path:hsa04010	*CACNA1I; CACNB3; TAB1; PRKCG; CACNA1G; MKNK2; DUSP7; DUSP16; RAF1; CACNA1B*
0,012470337	0,102233613	Circadian entrainment - Homo sapiens (human)	KEGG	path:hsa04713	*ADCY5; ADCY6; CACNA1G; PRKCG; CACNA1I*
0,013495836	0,102233613	Cortisol synthesis and secretion - Homo sapiens (human)	KEGG	path:hsa04927	*CACNA1I; ADCY6; ADCY5; CACNA1G*
0,014101188	0,102233613	Inflammatory mediator regulation of TRP channels - Homo sapiens (human)	KEGG	path:hsa04750	*ADCY5; ADCY6; ASIC1; PIK3R2; PRKCG*
0,014101188	0,102233613	Progesterone-mediated oocyte maturation - Homo sapiens (human)	KEGG	path:hsa04914	*ADCY5; ADCY6; FZR1; RAF1; PIK3R2*
0,015244563	0,105260078	Signaling pathways regulating pluripotency of stem cells - Homo sapiens (human)	KEGG	path:hsa04550	*ACVR2B; SMAD3; DVL3; PIK3R2; KAT6A; RAF1*
0,01784843	0,117637381	Hepatitis B - Homo sapiens (human)	KEGG	path:hsa05161	*PRKCG; SMAD3; PIK3R2; EP300; RAF1; CREBBP*
0,028325887	0,171388763	Type II diabetes mellitus - Homo sapiens (human)	KEGG	path:hsa04930	*CACNA1B; PIK3R2; CACNA1G*
0,028693946	0,171388763	Neurotrophin signaling pathway - Homo sapiens (human)	KEGG	path:hsa04722	*RAPGEF1; SH2B1; RAF1; PIK3R2; NTRK3*

0,029549787	0,171388763 Rap1 signaling pathway - Homo sapiens (human)	KEGG	path:hsa04015	*ADCY5; ADCY6; PRKCG; PIK3R2; SIPA1L1; RAF1; RAPGEF1*
0,033028417	0,176537577 Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate - Homo sapiens (human)	KEGG	path:hsa00532	*CHPF; CSGALNACT1*
0,033430507	0,176537577 Cell cycle - Homo sapiens (human)	KEGG	path:hsa04110	*EP300; FZR1; TFDP2; CREBBP; SMAD3*
0,034158859	0,176537577 TGF-beta signaling pathway - Homo sapiens (human)	KEGG	path:hsa04350	*ACVR2B; EP300; CREBBP; SMAD3*
0,038074295	0,176537577 GABAergic synapse - Homo sapiens (human)	KEGG	path:hsa04727	*ADCY5; ADCY6; CACNA1B; PRKCG*
0,038074295	0,176537577 Gap junction - Homo sapiens (human)	KEGG	path:hsa04540	*ADCY5; ADCY6; RAF1; PRKCG*
0,039434343	0,176537577 Longevity regulating pathway - Homo sapiens (human)	KEGG	path:hsa04211	*ADCY5; ADCY6; TSC2; PIK3R2*
0,039722595	0,176537577 Relaxin signaling pathway - Homo sapiens (human)	KEGG	path:hsa04926	*ADCY5; ADCY6; RAF1; PIK3R2; SMAD3*
0,040929812	0,176537577 Axon guidance - Homo sapiens (human)	KEGG	path:hsa04360	*SEMA3D; PLXNA1; PIK3R2; SSH1; SEMA4A; RAF1*
0,042236892	0,176537577 Morphine addiction - Homo sapiens (human)	KEGG	path:hsa05032	*ADCY5; ADCY6; CACNA1B; PRKCG*
0,042612519	0,176537577 Regulation of lipolysis in adipocytes - Homo sapiens (human)	KEGG	path:hsa04923	*ADCY5; ADCY6; PIK3R2*
0,047918025	0,193003155 Insulin signaling pathway - Homo sapiens (human)	KEGG	path:hsa04910	*MKNK2; RAF1; TSC2; PIK3R2; RAPGEF1*

p-value	q-value	pathway	source	external_id	members_input_overlap
0,006460603	0,295828262	Biosynthesis of unsaturated fatty acids - Homo sapiens (human)	KEGG	path:hsa01040	*HSD17B12; ACAA1; ACOT2*
0,006573961	0,295828262	Peroxisome - Homo sapiens (human)	KEGG	path:hsa04146	*CRAT; ECH1; ACAA1; ABCD4; IDH2*
0,012454505	0,373635144	Glycerophospholipid metabolism - Homo sapiens (human)	KEGG	path:hsa00564	*MBOAT2; AGPAT5; PCYT1B; PNPLA7; GPD2*
0,025033788	0,432178072	Thyroid hormone signaling pathway - Homo sapiens (human)	KEGG	path:hsa04919	*AKT1; PFKFB2; KAT2A; HIF1A; SIN3A*
0,030759886	0,432178072	Valine, leucine and isoleucine degradation - Homo sapiens (human)	KEGG	path:hsa00280	*IVD; ACAA1; BCKDHA*
0,032432318	0,432178072	Arginine and proline metabolism - Homo sapiens (human)	KEGG	path:hsa00330	*PYCR2; LAP3; MAOB*
0,045336727	0,432178072	Glycosaminoglycan biosynthesis - heparan sulfate / heparin - Homo sapiens (human)	KEGG	path:hsa00534	*NDST1; HS2ST1*
0,045432578	0,432178072	Glutathione metabolism - Homo sapiens (human)	KEGG	path:hsa00480	*IDH2; LAP3; GSTM3*
0,048828012	0,432178072	Glycosylphosphatidylinositol (GPI)-anchor biosynthesis - Homo sapiens (human)	KEGG	path:hsa00563	*PIGP; PIGS*
0,049554342	0,432178072	Endometrial cancer - Homo sapiens (human)	KEGG	path:hsa05213	*APC2; AKT1; CTNNA2*

p-value	q-value	pathway	source	external_id	members_input_overlap
0,000108699	0,01695621	ErbB signaling pathway - Homo sapiens (human)	KEGG	path:hsa04012	*MAPK3; PRKCA; SHC3; CAMK2A; AKT1; BAD; PAK6*
0,000318575	0,01695621	Hepatocellular carcinoma - Homo sapiens (human)	KEGG	path:hsa05225	*DPF1; SMARCB1; SHC3; PRKCA; AKT1; BAD; APC; SMARCA4; MAPK3*
0,000386328	0,01695621	Insulin signaling pathway - Homo sapiens (human)	KEGG	path:hsa04910	*RPTOR; SOCS4; SHC3; PTPRF; BAD; FASN; AKT1; MAPK3*
0,000429271	0,01695621	MAPK signaling pathway - Homo sapiens (human)	KEGG	path:hsa04010	*RPS6KA4; MAPK3; PRKCA; MRAS; JUND; AKT1; MAPK11; ARRB1; SRF; MAP3K5; MAPK8IP1; TAOK3*
0,000753484	0,021319832	Thermogenesis - Homo sapiens (human)	KEGG	path:hsa04714	*RPTOR; DPF1; SMARCB1; MGLL; MAP3K5; COX20; MAPK11; KDM3B; KDM3A; SMARCA4*
0,000857067	0,021319832	Neurotrophin signaling pathway - Homo sapiens (human)	KEGG	path:hsa04722	*MAPK3; SHC3; CAMK2A; MAP3K5; AKT1; BAD; MAPK11*
0,00094455	0,021319832	VEGF signaling pathway - Homo sapiens (human)	KEGG	path:hsa04370	*PRKCA; AKT1; BAD; MAPK11; MAPK3*
0,001080076	0,021331508	Focal adhesion - Homo sapiens (human)	KEGG	path:hsa04510	*SHC3; PRKCA; CCND3; TNR; AKT1; BAD; PAK6; PARVA; MAPK3*
0,001437178	0,025230465	Relaxin signaling pathway - Homo sapiens (human)	KEGG	path:hsa04926	*MAPK3; PRKCA; SHC3; GNB2; AKT1; MAPK11; ARRB1*
0,001912998	0,028586817	Renal cell carcinoma - Homo sapiens (human)	KEGG	path:hsa05211	*AKT1; BAD; PAK6; TFE3; MAPK3*
0,002039091	0,028586817	Prolactin signaling pathway - Homo sapiens (human)	KEGG	path:hsa04917	*SOCS4; SHC3; MAPK11; MAPK3; AKT1*
0,002171151	0,028586817	Glioma - Homo sapiens (human)	KEGG	path:hsa05214	*PRKCA; AKT1; SHC3; CAMK2A; MAPK3*
0,003233772	0,03930277	mTOR signaling pathway - Homo sapiens (human)	KEGG	path:hsa04150	*RPTOR; MAPK3; PRKCA; STK11; MAPKAP1; AKT1; CLIP1*
0,00381969	0,043107932	Thyroid hormone signaling pathway - Homo sapiens (human)	KEGG	path:hsa04919	*MAPK3; PRKCA; THRA; AKT1; BAD; SIN3A*
0,004738684	0,049914135	Insulin secretion - Homo sapiens (human)	KEGG	path:hsa04911	*RAPGEF4; PRKCA; CAMK2A; RIMS2; KCNN2*
0,006159059	0,057185862	Autophagy - animal - Homo sapiens (human)	KEGG	path:hsa04140	*RPTOR; MAPK3; STK11; MRAS; AKT1; BAD*
0,006318747	0,057185862	Morphine addiction - Homo sapiens (human)	KEGG	path:hsa05032	*PRKCA; GABRB1; ARRB1; PDE7B; GNB2*
0,006514845	0,057185862	Endometrial cancer - Homo sapiens (human)	KEGG	path:hsa05213	*APC; BAD; MAPK3; AKT1*
0,007198517	0,058846891	Galactose metabolism - Homo sapiens (human)	KEGG	path:hsa00052	*B4GALT2; PFKL; PGM1*
0,007448974	0,058846891	Axon guidance - Homo sapiens (human)	KEGG	path:hsa04360	*EPHB1; MAPK3; PRKCA; CAMK2A; ROBO1; PAK6; SEMA6B*
0,009332785	0,065020266	HIF-1 signaling pathway - Homo sapiens (human)	KEGG	path:hsa04066	*PRKCA; AKT1; PFKL; CAMK2A; MAPK3*
0,010227281	0,065020266	Non-small cell lung cancer - Homo sapiens (human)	KEGG	path:hsa05223	*PRKCA; MAPK3; BAD; AKT1*

0,010251854	0,065020266	Ras signaling pathway - Homo sapiens (human)	KEGG	path:hsa04014	*SHC3; PRKCA; MRAS; GNB2; AKT1; BAD; PAK6; MAPK3*
0,010712101	0,065020266	Adrenergic signaling in cardiomyocytes - Homo sapiens (human)	KEGG	path:hsa04261	*MAPK3; PRKCA; CAMK2A; RAPGEF4; AKT1; MAPK11*

0,010770089	0,065020266	Fc epsilon RI signaling pathway - Homo sapiens (human)	KEGG	path:hsa04664	PRKCA; MAPK3; MAPK11; AKT1
0,011058958	0,065020266	Phospholipase D signaling pathway - Homo sapiens (human)	KEGG	path:hsa04072	MAPK3; PRKCA; SHC3; MRAS; RAPGEF4; AKT1
0,011111058	0,065020266	Chemokine signaling pathway - Homo sapiens (human)	KEGG	path:hsa04062	MAPK3; SHC3; GNB2; AKT1; BAD; ARRB1; CXCL14
0,011818728	0,066181324	Parathyroid hormone synthesis, secretion and action - Homo sapiens (human)	KEGG	path:hsa04928	PRKCA; MMP24; MAPK3; JUND; ARRB1
0,012147205	0,066181324	Retrograde endocannabinoid signaling - Homo sapiens (human)	KEGG	path:hsa04723	PRKCA; MGLL; GNB2; MAPK3; MAPK11; GABRB1
0,013705595	0,07014442	TNF signaling pathway - Homo sapiens (human)	KEGG	path:hsa04668	AKT1; RPS6KA4; MAPK3; MAPK11; MAP3K5
0,013762513	0,07014442	Adherens junction - Homo sapiens (human)	KEGG	path:hsa04520	IQGAP1; PTPRF; MAPK3; WASF3
0,014721028	0,070854532	Cholinergic synapse - Homo sapiens (human)	KEGG	path:hsa04725	PRKCA; AKT1; MAPK3; CAMK2A; GNB2
0,015166966	0,070854532	Proteoglycans in cancer - Homo sapiens (human)	KEGG	path:hsa05205	MAPK3; PRKCA; CAMK2A; MRAS; AKT1; MAPK11; IQGAP1
0,015247178	0,070854532	Serotonergic synapse - Homo sapiens (human)	KEGG	path:hsa04726	PRKCA; GABRB1; MAPK3; KCNN2; GNB2
0,0158706	0,071644423	PI3K-Akt signaling pathway - Homo sapiens (human)	KEGG	path:hsa04151	RPTOR; TNR; PRKCA; CCND3; STK11; GNB2; AKT1; BAD; MAGI2; MAPK3
0,016502337	0,07175408	Chronic myeloid leukemia - Homo sapiens (human)	KEGG	path:hsa05220	AKT1; BAD; SHC3; MAPK3
0,017134372	0,07175408	Rap1 signaling pathway - Homo sapiens (human)	KEGG	path:hsa04015	MAPK3; PRKCA; MRAS; RAPGEF4; AKT1; MAGI2; MAPK11
0,01725731	0,07175408	Cellular senescence - Homo sapiens (human)	KEGG	path:hsa04218	MAPK3; CCND3; MRAS; AKT1; HIPK1; MAPK11
0,018066937	0,073194258	Sphingolipid signaling pathway - Homo sapiens (human)	KEGG	path:hsa04071	PRKCA; AKT1; MAPK11; MAPK3; MAP3K5
0,019285124	0,07601647	AMPK signaling pathway - Homo sapiens (human)	KEGG	path:hsa04152	RPTOR; AKT1; PFKL; FASN; STK11
0,019725793	0,07601647	Human immunodeficiency virus 1 infection - Homo sapiens (human)	KEGG	path:hsa05170	MAPK3; PRKCA; GNB2; AKT1; BAD; MAPK11; PAK6
0,02207144	0,083030655	Peroxisome - Homo sapiens (human)	KEGG	path:hsa04146	CRAT; PEX10; ACOX1; PMVK
0,024769473	0,089613885	Colorectal cancer - Homo sapiens (human)	KEGG	path:hsa05210	AKT1; BAD; MAPK3; APC
0,024955765	0,089613885	N-Glycan biosynthesis - Homo sapiens (human)	KEGG	path:hsa00510	B4GALT2; ST6GAL2; MGAT4C
0,026674116	0,091150704	GABAergic synapse - Homo sapiens (human)	KEGG	path:hsa04727	PRKCA; GABRB1; SLC6A1; GNB2
0,026950191	0,091150704	Dopaminergic synapse - Homo sapiens (human)	KEGG	path:hsa04728	PRKCA; AKT1; MAPK11; CAMK2A; GNB2
0,027658489	0,091150704	Longevity regulating pathway - Homo sapiens (human)	KEGG	path:hsa04211	RPTOR; EHMT2; STK11; AKT1
0,027691353	0,091150704	Amyotrophic lateral sclerosis (ALS) - Homo sapiens (human)	KEGG	path:hsa05014	BAD; MAPK11; MAP3K5
0,028664333	0,09242785	Fc gamma R-mediated phagocytosis - Homo sapiens (human)	KEGG	path:hsa04666	PRKCA; AKT1; MAPK3; WASF3

0,031811346	0,095563634	GnRH signaling pathway - Homo sapiens (human)	KEGG	path:hsa04912	PRKCA; MAPK3; CAMK2A; MAPK11
0,031811346	0,095563634	IL-17 signaling pathway - Homo sapiens (human)	KEGG	path:hsa04657	MAPK3; JUND; MAPK4; MAPK11
0,031846889	0,095563634	Apelin signaling pathway - Homo sapiens (human)	KEGG	path:hsa04371	AKT1; MAPK3; HDAC5; MRAS; GNB2
0,032056156	0,095563634	Other types of O-glycan biosynthesis - Homo sapiens (human)	KEGG	path:hsa00514	B4GALT2; ST6GAL2
0,034817961	0,100986999	Mannose type O-glycan biosynthesis - Homo sapiens (human)	KEGG	path:hsa00515	ISPD; B4GALT2
0,035153702	0,100986999	Circadian entrainment - Homo sapiens (human)	KEGG	path:hsa04713	PRKCA; MAPK3; CAMK2A; GNB2
0,038692426	0,107252689	AGE-RAGE signaling pathway in diabetic complications - Homo sapiens (human)	KEGG	path:hsa04933	PRKCA; AKT1; MAPK11; MAPK3
0,038692426	0,107252689	Choline metabolism in cancer - Homo sapiens (human)	KEGG	path:hsa05231	PRKCA; AKT1; MAPK3; WASF3
0,040156863	0,109392834	Lysine degradation - Homo sapiens (human)	KEGG	path:hsa00310	EHMT2; SETD2; HADHA
0,041160953	0,110227636	T cell receptor signaling pathway - Homo sapiens (human)	KEGG	path:hsa04660	AKT1; PAK6; MAPK11; MAPK3
0,043560496	0,114709306	cAMP signaling pathway - Homo sapiens (human)	KEGG	path:hsa04024	MAPK3; ACOX1; CAMK2A; RAPGEF4; AKT1; BAD
0,045027918	0,115572443	C-type lectin receptor signaling pathway - Homo sapiens (human)	KEGG	path:hsa04625	AKT1; MAPK11; MRAS; MAPK3
0,045351212	0,115572443	Viral carcinogenesis - Homo sapiens (human)	KEGG	path:hsa05203	HDAC5; CCND3; MAPK3; BAD; HDAC11; SRF

p-value	q-value	pathway	source	external_id	members_input_overlap
1,78E-07	4,26E-06	Allograft rejection - Homo sapiens (human)	KEGG	path:hsa05330	*HLA-C; HLA-B; HLA-A; HLA-DRB1; HLA-E*
2,76E-07	4,26E-06	Antigen processing and presentation - Homo sapiens (human)	KEGG	path:hsa04612	*HLA-C; HLA-B; HLA-A; HLA-DRB1; HLA-E; B2M*
3,03E-07	4,26E-06	Graft-versus-host disease - Homo sapiens (human)	KEGG	path:hsa05332	*HLA-C; HLA-B; HLA-A; HLA-DRB1; HLA-E*
3,87E-07	4,26E-06	Type I diabetes mellitus - Homo sapiens (human)	KEGG	path:hsa04940	*HLA-C; HLA-B; HLA-A; HLA-DRB1; HLA-E*
9,12E-07	7,49E-06	Phagosome - Homo sapiens (human)	KEGG	path:hsa04145	*HLA-C; HLA-B; HLA-A; HLA-DRB1; HLA-E; CD14; C1R*
1,02E-06	7,49E-06	Autoimmune thyroid disease - Homo sapiens (human)	KEGG	path:hsa05320	*HLA-C; HLA-B; HLA-A; HLA-DRB1; HLA-E*
1,77E-06	1,11E-05	Viral myocarditis - Homo sapiens (human)	KEGG	path:hsa05416	*HLA-C; HLA-B; HLA-A; HLA-DRB1; HLA-E*
5,93E-06	3,26E-05	Epstein-Barr virus infection - Homo sapiens (human)	KEGG	path:hsa05169	*STAT3; HLA-C; HLA-B; HLA-A; HLA-DRB1; HLA-E; B2M*
8,70E-06	4,25E-05	Human immunodeficiency virus 1 infection - Homo sapiens (human)	KEGG	path:hsa05170	*SAMHD1; HLA-C; HLA-B; HLA-A; HLA-E; B2M; BST2*
4,48E-05	0,000184828	Herpes simplex infection - Homo sapiens (human)	KEGG	path:hsa05168	*HLA-C; HLA-B; HLA-A; HLA-DRB1; HLA-E; PML*
4,62E-05	0,000184828	Kaposi sarcoma-associated herpesvirus infection - Homo sapiens (human)	KEGG	path:hsa05167	*STAT3; HLA-C; HLA-B; HLA-A; HLA-E; ANGPT2*
0,00011421	0,000418788	Human T-cell leukemia virus 1 infection - Homo sapiens (human)	KEGG	path:hsa05166	*HLA-C; HLA-B; HLA-A; HLA-DRB1; HLA-E; B2M*
0,00013245	0,000448301	Human cytomegalovirus infection - Homo sapiens (human)	KEGG	path:hsa05163	*STAT3; HLA-C; HLA-B; HLA-A; HLA-E; B2M*
0,00014355	0,000451149	Cell adhesion molecules (CAMs) - Homo sapiens (human)	KEGG	path:hsa04514	*HLA-C; HLA-B; HLA-A; HLA-DRB1; HLA-E*
0,00069028	0,002024833	Viral carcinogenesis - Homo sapiens (human)	KEGG	path:hsa05203	*HLA-C; HLA-B; HLA-A; STAT3; HLA-E*
0,00112791	0,003101764	Natural killer cell mediated cytotoxicity - Homo sapiens (human)	KEGG	path:hsa04650	*HLA-C; HLA-B; HLA-A; HLA-E*
0,00162591	0,004008823	Acute myeloid leukemia - Homo sapiens (human)	KEGG	path:hsa05221	*PML; CD14; STAT3*
0,00163997	0,004008823	Endocytosis - Homo sapiens (human)	KEGG	path:hsa04144	*HLA-C; HLA-B; HLA-A; HLA-E; PML*
0,0024171	0,005361313	Cellular senescence - Homo sapiens (human)	KEGG	path:hsa04218	*HLA-C; HLA-B; HLA-A; HLA-E*
0,00243696	0,005361313	Pertussis - Homo sapiens (human)	KEGG	path:hsa05133	*C1R; CD14; SERPING1*
0,0161655	0,033870562	Staphylococcus aureus infection - Homo sapiens (human)	KEGG	path:hsa05150	*C1R; HLA-DRB1*
0,02214313	0,042364786	Inflammatory bowel disease (IBD) - Homo sapiens (human)	KEGG	path:hsa05321	*HLA-DRB1; STAT3*
0,02214523	0,042364786	Tight junction - Homo sapiens (human)	KEGG	path:hsa04530	*MSN; MYH9; MYL12A*
0,02475343	0,045381295	Adipocytokine signaling pathway - Homo sapiens (human)	KEGG	path:hsa04920	*STAT3; ACSL5*
0,03145622	0,053805206	Human papillomavirus infection - Homo sapiens (human)	KEGG	path:hsa05165	*HLA-C; HLA-B; HLA-A; HLA-E*
0,03179399	0,053805206	Complement and coagulation cascades - Homo sapiens (human)	KEGG	path:hsa04610	*C1R; SERPING1*
0,03713614	0,060518155	Salmonella infection - Homo sapiens (human)	KEGG	path:hsa05132	*CD14; MYH9*

0,0393766	0,061877507	Regulation of actin cytoskeleton - Homo sapiens (human)	KEGG	path:hsa04810	*MSN; MYH9; MYL12A*
0,04531691	0,068756692	Hematopoietic cell lineage - Homo sapiens (human)	KEGG	path:hsa04640	*CD14; HLA-DRB1*
0,04875978	0,07151435	HIF-1 signaling pathway - Homo sapiens (human)	KEGG	path:hsa04066	*ANGPT2; STAT3*

p-value	q-value	pathway	source	external_id	members_input_overlap
4,41E-05	0,000671367	NOD-like receptor signaling pathway - Homo sapiens (human)	KEGG	path:hsa04621	*OAS3; NAMPT; STAT2; STAT1; GBP4*
4,80E-05	0,000671367	Influenza A - Homo sapiens (human)	KEGG	path:hsa05164	*TRIM25; OAS3; STAT2; STAT1; MX1*
0,000100845	0,000941224	Epstein-Barr virus infection - Homo sapiens (human)	KEGG	path:hsa05169	*TAP1; GADD45A; OAS3; STAT2; STAT1*
0,000251448	0,001760136	Measles - Homo sapiens (human)	KEGG	path:hsa05162	*OAS3; STAT2; STAT1; MX1*
0,000476972	0,002671046	Hepatitis C - Homo sapiens (human)	KEGG	path:hsa05160	*OAS3; STAT2; STAT1; MX1*
0,000926754	0,004324853	Herpes simplex infection - Homo sapiens (human)	KEGG	path:hsa05168	*TAP1; OAS3; STAT2; STAT1*
0,006620327	0,025553451	JAK-STAT signaling pathway - Homo sapiens (human)	KEGG	path:hsa04630	*STAT2; STAT1; PIM1*
0,00757761	0,025553451	Pathways in cancer - Homo sapiens (human)	KEGG	path:hsa05200	*ZBTB16; GADD45A; STAT2; STAT1; PIM1*
0,008213609	0,025553451	Human papillomavirus infection - Homo sapiens (human)	KEGG	path:hsa05165	*COL9A2; STAT2; STAT1; MX1*
0,01009434	0,026264841	Chemokine signaling pathway - Homo sapiens (human)	KEGG	path:hsa04062	*STAT2; STAT1; SHC1*
0,010747314	0,026264841	Acute myeloid leukemia - Homo sapiens (human)	KEGG	path:hsa05221	*ZBTB16; PIM1*
0,012028873	0,026264841	Prolactin signaling pathway - Homo sapiens (human)	KEGG	path:hsa04917	*SHC1; STAT1*
0,012359281	0,026264841	Glioma - Homo sapiens (human)	KEGG	path:hsa05214	*GADD45A; SHC1*
0,013720397	0,026264841	Pancreatic cancer - Homo sapiens (human)	KEGG	path:hsa05212	*GADD45A; STAT1*
0,014070451	0,026264841	Chronic myeloid leukemia - Homo sapiens (human)	KEGG	path:hsa05220	*GADD45A; SHC1*
0,023151869	0,04051577	AGE-RAGE signaling pathway in diabetic complications - Homo sapiens (human)	KEGG	path:hsa04933	*STAT1; PIM1*
0,025374886	0,041793931	C-type lectin receptor signaling pathway - Homo sapiens (human)	KEGG	path:hsa04625	*STAT2; STAT1*
0,03664858	0,057008903	Osteoclast differentiation - Homo sapiens (human)	KEGG	path:hsa04380	*STAT2; STAT1*
0,046002736	0,065208033	Hepatitis B - Homo sapiens (human)	KEGG	path:hsa05161	*STAT2; STAT1*
0,047736722	0,065208033	Breast cancer - Homo sapiens (human)	KEGG	path:hsa05224	*GADD45A; SHC1*
0,048906025	0,065208033	Gastric cancer - Homo sapiens (human)	KEGG	path:hsa05226	*GADD45A; SHC1*

p-value	q-value	pathway	source	external_id	members_input_overlap
0,003726703	0,040443648	Ras signaling pathway - Homo sapiens (human)	KEGG	path:hsa04014	*HGF; PLA2G16; GAB2; FOXO4*
0,008000614	0,040443648	Gastric cancer - Homo sapiens (human)	KEGG	path:hsa05226	*HGF; TGFB3; CTNNA3*
0,00808873	0,040443648	Malaria - Homo sapiens (human)	KEGG	path:hsa05144	*HGF; TGFB3*
0,011090007	0,041587527	Hepatocellular carcinoma - Homo sapiens (human)	KEGG	path:hsa05225	*HGF; SMARCC1; TGFB3*
0,0155874	0,0467622	Renal cell carcinoma - Homo sapiens (human)	KEGG	path:hsa05211	*HGF; TGFB3*
0,018708677	0,046771693	Chronic myeloid leukemia - Homo sapiens (human)	KEGG	path:hsa05220	*TGFB3; GAB2*
0,046098288	0,083026448	Cell cycle - Homo sapiens (human)	KEGG	path:hsa04110	*MCM7; TGFB3*
0,047694745	0,083026448	MAPK signaling pathway - Homo sapiens (human)	KEGG	path:hsa04010	*HGF; TGFB3; MAP4K4*

A.2 Module networks-cytoscape

- Control black
- Case black
- Control Pink
- Case Pink

Gene name	BetweennessCentrality	Degree
CREBBP	1.3505E-4	10
CHERP	0.00282056	25
EP300	0.00102706	12
TBC1D13	0.00297901	28
GIT1	0.00369593	17
ZCCHC14	0.03206978	41
CELF3	0.0069203	26
SMAD3	0.00786357	22
KMT2D	6.0762E-4	19
ZNF768	0.0016751	20
EP400	0.00334982	23
ITGA3	0.00174144	12
PPP1R37	0.00415387	46
RAPGEFL1	0.00315502	22
ZNF142	0.00186673	19
VPS13B	0.01345113	40
CACNA1B	8.9311E-4	14
FOXK1	0.0321876	43
SH2B1	0.00501494	22
CSMD1	0.00174431	32
ZNF512B	0.01171222	20
KIAA0753	8.963E-5	5
CACNA1G	0.01091996	36
KCNQ2	0.00582876	53
BRPF3	0.02572826	75
MAST3	0.0565805	92
CACNA1I	8.733E-5	8
SGSM3	9.6312E-4	16
NUMBL	0.01702674	67
PIK3R2	0.00263171	47
SP4	0.03621824	22
PLXNA1	0.01464119	45
SCAP	0.01895989	71
IQSEC2	0.07597296	97
CPNE5	0.010769	63
ABHD8	9.3359E-4	35
VHL	1.001E-5	6
RBM26	0.03915563	88
NCOR1	0.00965873	61
ASXL3	0.0018421	36
FOXK2	0.03468027	86
TTBK1	0.03079873	83
FBXO10	0.0	5
SORBS2	0.00307857	33
NBPF8	0.0	5
PYGO2	0.00337759	37
DUSP7	0.00472252	28
SUN1	0.00290118	21
CACNB3	0.00766593	56
SNTB2	0.02168873	79
GPRIN1	0.0077486	59
GPR137	8.8373E-4	29
ANGEL2	0.00773542	14
AATK	8.424E-5	8
TNFAIP8L1	9.7113E-4	18
ELAVL3	0.02013881	72
ZNF667	0.00923466	28
DGCR5	0.0030178	45
ZNF585B	0.00255453	43
MARK4	0.02177733	56
CDK13	0.00291018	46
NAV3	0.02244846	70
LPHN1	0.04351122	65
RAPGEF1	0.00740547	6
DUSP16	0.0	10
XRN1	0.00594322	13
ARID1A	0.00565215	25
CASP8AP2	0.00285759	32
BAI2	0.03018519	66
MED13L	0.00913655	26
METTL8	0.02913422	51
INO80	0.00702772	38
MPRIP	0.01232049	62
NTRK3	0.0	3
PLEKHA6	0.01125006	33
SETBP1	0.00284025	23
SCAF4	0.00279596	17
SHANK2	0.01045958	30
SLC30A7	0.0072796	60
BSN	0.02039259	73
BRD3	0.06845741	59
NLGN2	0.01468089	53
ATF7IP	0.0016123	15
CHD2	0.0035966	8
ADCY6	4.0108E-4	13
BBS1	5.2216E-4	10
LRRN1	0.01868698	65
KIAA2018	0.02443237	49
SLC25A22	0.00316088	47
SLC26A11	0.00751349	57
KIAA2026	0.02422422	78
SLC35E2B	0.01985298	51
KIAA1671	0.01338998	54
SYNGAP1	0.07263409	94
ZNF43	0.00811723	13
ATXN2	4.2388E-4	17
ERCC6	0.04635542	61
RP11-252A24.7	0.00128221	26

CYTH3	0.00729927	2
ZNF786	0.0	1
SPRN	0.02075982	23
CELSR3	0.00657371	19
NFX1	0.01019995	21
IQCE	0.0041824	15
ZNF236	0.01089249	22
AGAP2	1.8614E-4	7
KLHL36	0.00681982	20
PKNOX2	1.4187E-4	4
DMWD	0.00486166	27
CTSO	5.6544E-4	7
NSMAF	0.00216297	6
MAST2	0.0038561	24
ZNF382	2.6247E-4	3
ARID2	0.00865529	11
ZNF585A	1.2675E-4	3
GET4	0.00593777	29
CYP46A1	5.8674E-4	6
CBFA2T2	0.01355397	16
PIANP	1.681E-5	6
ZNF678	6.998E-4	7
GTF2H2C	8.9843E-4	4
SHISA7	0.02878993	36
USP2	0.00117363	36
R3HDM1	0.0010985	18
MOXD1	3.82E-6	8
JPH1	0.00427542	46
KIAA1377	6.0548E-4	23
BCL7A	0.00289394	50
SIN3B	0.0146752	18
PHIP	1.3553E-4	30
ZNF772	0.00188117	31
RP5-837J1.2	0.00665523	21
FAM65A	0.00288785	5
TSPAN14	7.2322E-4	6
TFDP2	0.00352405	9
SNX21	0.00505899	19
PEAK1	0.02689746	18
C20orf112	0.01030944	20
PTPN14	2.81E-5	6
CNIH2	0.01347127	30
CHST1	1.5074E-4	16
SLC35E2	0.00122307	25
SUGP2	3.2434E-4	9
ZNF318	0.00824866	8
SPEN	5.6783E-4	5
UBR5	0.0054048	7
ERGIC1	0.00570404	10
NIPBL	5.6783E-4	5

KAT6A	1.9659E-4	27
PPP6R1	6.458E-5	26
AGAP3	7.9968E-4	34
SIPA1L1	1.3993E-4	17
C19orf26	5.1122E-4	22
FOXRED2	0.00265147	14
RNF24	0.00337246	30
SMPD3	5.9015E-4	22
HOMER2	0.00650147	17
CHPF	2.1601E-4	19
PRKCG	1.1785E-4	16
PTCD3	2.0931E-4	15
CLUH	0.00275188	17
DAGLA	0.00317361	32
TMEM8B	0.00110264	24
MINK1	5.4765E-4	17
ZNF687	9.5361E-4	19
RNF44	0.00382545	21
ICA1L	0.00210569	24
GRIK2	0.00162131	13
PACS1	4.661E-5	10
TOM1L2	1.492E-4	17
KIAA0895L	0.00433549	29
DCHS2	0.00457151	25
TMEM184B	0.00115034	20
SLC35F6	0.00826533	22
NDST1	1.432E-5	3
LMF2	0.00108516	5
TTYH3	0.00113573	4
SBK1	0.00628783	7
ASIC1	0.00593127	16
NACAD	0.01033167	24
DISP2	5.2318E-4	20
NPAS2	0.01264224	12
BRSK2	0.01125661	14
MPDZ	2.2373E-4	25
CNOT6	4.843E-5	25
OPRL1	2.119E-5	13
ARL10	0.01012902	25
ZNF445	0.0089005	17
PTPN23	0.00184942	21
DLGAP4	0.00171851	14
SPG11	0.00362589	19
KCNC4	7.299E-5	8
KMT2A	6.8044E-4	13
LATS1	0.00110632	20
ANKFY1	0.0037605	10
RTN4RL2	8.6433E-4	12
SEMA4A	0.00801836	9
TRIM9	0.0	3

DVL3	6.3765E-4	5
SSH1	5.69E-4	5
GOLGA3	0.00865222	10
LRRC28	0.00156154	5
BAIAP2	0.00109193	7
ZNF808	0.00156154	5
NCOA6	0.01334041	14
ADCY5	0.01570492	15
MFN1	0.00160787	9
POM121	0.00197161	11
SPTBN4	0.0	3
FUS	0.00553931	13
TNRC6B	0.00700821	20
STX16	4.247E-5	9
MTO1	0.0	3
C11orf30	7.0889E-4	3
ZNF568	0.00827062	21
LRIG2	0.02122579	16
TUG1	0.00154978	19
RAF1	8.3704E-4	6
ZBTB34	8.82E-5	5
NELFCD	2.8375E-4	12
USP43	0.0	3
RNF169	2.105E-5	20
BTBD9	0.0073015	7
LSM14A	1.1461E-4	17
CLIP2	4.396E-4	14
ARHGAP21	2.571E-5	10
DPYSL3	0.0	2
RAB40B	5.5124E-4	19
GRM2	1.2965E-4	11
ZNF704	0.01190445	14
SGK223	0.0	8
INF2	4.7489E-4	6
SNAI3-AS1	0.0	3
BDP1	0.00137378	11
PSMG3-AS1	0.0	7
MKNK2	8.294E-5	7
HIST2H2BE	0.0	4
RASD2	0.0	1
TAB1	8.943E-5	3
PHF10	0.0	2
RBM25	1.9453E-4	11
PHC3	1.213E-4	13
TSC2	9.7606E-4	7
HIST2H2AA4	7.091E-5	5
DFFA	0.02892971	5
MYO18A	0.0	1
TTLL11	5.2705E-4	6
ZBTB10	1.3309E-4	6

DSCR3	0.00788002	6
ZNF708	0.0	4
MDGA1	0.0	1
KANSL3	1.7223E-4	3
NAV1	0.0	3
ACVR2B	1.3144E-4	6
SRRM1	6.032E-5	2
ASXL2	0.0139217	7
GPR83	0.0	2
LRRC6	1.337E-5	3
TRERF1	0.00729927	2
DENND4B	0.0	1
NCLN	0.0	1
PTPN12	0.0	1
PPP1R1A	1.337E-5	3
SLC39A11	0.00363627	4
FER	0.00363627	4
SGCD	0.0	2
PRRC1	0.0	2
PIGM	0.0	5
TMEM108	0.00440361	6
TOR1AIP2	0.0	3
ZNF100	3.124E-5	4
SHISA9	0.00988878	6
CSGALNACT1	0.0	2
SEMA3D	0.0	1
NDUFA6-AS1	0.0	1

Gene name	BetweennessCentrality	Degree
LAS1L	0.0	4
SEC63	0.27990696	295
ADSS	0.02472642	197
POLR3E	0.17865454	361
KLHL22	0.1928226	365
CYP51A1	0.0	3
SNAP91	0.1119059	223
FSCN1	0.15246511	325
BAD	0.0	4
MRTO4	0.00797003	57
ADRBK2	0.03819192	216
MAD1L1	0.0	3
M6PR	0.0	3
ICA1	0.0	1
MTMR7	2.7198E-4	14
SARM1	0.01466781	168
SAMD4A	0.02266278	123
LSG1	0.00111071	38
EML1	0.07027997	207
TGFBR3	0.00231992	61
NUAK1	0.02351188	182
IPCEF1	0.03551433	202
DDX24	0.00195341	66
ACOT7	0.01514592	186
SLC7A2	0.0	4
ARF5	2.151E-4	13
BID	0.00307421	82
CAMKK1	2.1299E-4	12
DHX33	8.68E-6	12
PRKAR2B	0.0010524	64
CREBBP	2.5854E-4	21
IBTK	0.0	8
PDK2	8.19E-6	12
GGNBP2	0.0	9
MAP3K9	7.126E-5	10
PHTF2	7.126E-5	10
KIAA0100	1.238E-5	12
MATK	3.2651E-4	17
LUC7L	1.952E-5	14
UBE3C	2.4009E-4	16
RANBP9	6.98E-5	7
UQCRC1	0.0	7
HIVEP2	3.2127E-4	20
TSPAN9	2.5937E-4	22
PTBP1	4.4542E-4	15
RABGAP1	5.7538E-4	22
EXTL3	1.588E-5	14
NUB1	2.151E-4	12
RWDD2A	2.1564E-4	15
DNASE1L1	2.8229E-4	14
GPRC5A	5.173E-4	29
MATR3	1.4384E-4	11
STMN4	1.061E-5	11
GLT8D1	3.847E-4	22
ATP2C1	3.8201E-4	17
AGPS	2.22E-4	14
NRXN3	1.597E-5	12
FHL1	2.151E-4	12
GABRA1	1.5437E-4	14
NDUFS1	0.0	7
RB1CC1	3.7E-7	10
VEZT	7.126E-5	9
TBPL1	2.22E-4	14
BCLAF1	2.3073E-4	22
TSSC1	7.063E-5	10
UBA6	4.31E-6	9
ATP6V0A1	3.421E-4	19
APBA2	2.8563E-4	16
TMSB10	1.122E-5	9
TIMP2	3.7E-7	10
MAT2B	0.0	7
EDC4	0.0	8
STAU2	5.3336E-4	20
PSMA4	1.659E-5	14
POLR2B	0.0	6
XK	1.526E-5	11
GOPC	3.9122E-4	23
JKAMP	2.1299E-4	12
PIK3CB	5.9837E-4	25
NNAT	0.0	10
AP5M1	2.1526E-4	13
FAM168A	0.0	10
EIF2AK2	3.8709E-4	10
PUM2	3.3186E-4	17
C4orf27	1.0916E-4	16
IL17RB	2.1177E-4	13
PET112	1.5343E-4	15
STYK1	5.8609E-4	24
MPC1	1.2472E-4	33
GUCY1B3	0.0	9
MRPS24	5.6471E-4	21
VMP1	0.0	9
EIF4B	7.25E-6	10
MTMR1	2.0645E-4	10
HAGH	0.0	9
ZC3H15	2.3629E-4	20
SLK	4.7025E-4	38
ME1	0.0	8
SLC9A7	2.9569E-4	29

MSANTD3	0.0	5
ATG2B	0.0	9
CACNB1	2.0645E-4	11
DNTTIP2	2.159E-5	15
DHX8	1.061E-5	11
OTUD5	0.0	4
GRIPAP1	2.094E-5	16
IFT80	4.2061E-4	24
ERLEC1	1.559E-5	13
DNAJA2	2.3772E-4	20
RORA	2.8437E-4	16
ATP1B3	1.04E-6	12
ZXDC	0.0	11
PABPC1	2.8436E-4	16
MBD3	2.075E-4	15
TRNT1	4.0719E-4	12
ACADVL	2.153E-4	14
SIDT1	1.0883E-4	15
PPP2R2C	2.22E-4	15
RAB7A	2.034E-4	6
MCM6	2.8302E-4	14
ACTL6B	6.4E-7	7
RFX3	2.43E-6	8
RIF1	4.2492E-4	28
BZW1	2.534E-5	18
GEMIN5	2.22E-4	14
ULK2	8.0996E-4	21
CHMP2B	2.2403E-4	19
EIF3I	1.952E-5	14
SEH1L	1.497E-5	12
RRN3	3.8175E-4	21
POMGNT1	2.8302E-4	14
MAST2	5.3036E-4	21
EIF2AK1	1.108E-5	10
MRPL28	6.98E-5	8
ERO1LB	2.5953E-4	15
GNAS	2.22E-4	14
ERGIC2	1.03E-5	13
C3orf18	2.034E-4	8
CRLS1	2.22E-4	14
DYNLL1	7.7834E-4	22
TESC	7.386E-5	12
RPH3A	8.8183E-4	35
NOS1	5.6025E-4	20
CDIP1	5.1E-7	11
OSBPL8	1.064E-5	13
NLRP1	0.0	7
SCFD1	0.0	8
SUPT16H	0.00115384	56
SEC22C	0.0	8

DHPS	2.1447E-4	13
HSP90AB1	3.3038E-4	16
ITPR3	0.0	9
MKNK2	1.3665E-4	12
RANBP1	2.1592E-4	15
YPEL1	7.126E-5	12
PES1	3.9298E-4	17
MIEF1	2.1894E-4	18
KCNK10	7.063E-5	9
FKBP3	8.39E-6	12
ERH	1.2938E-4	19
SLC8A3	3.1549E-4	20
MTHFD1	3.0679E-4	17
PCNX	1.572E-5	13
GSKIP	0.0	9
PSMB5	7.167E-5	12
SRP54	3.0679E-4	17
DCAF11	2.22E-4	14
TRPC4AP	0.0	11
UQCC1	5.92E-6	15
PFDN4	4.087E-5	16
DOK5	2.2312E-4	17
PRPF6	7.475E-5	16
SEC23B	7.413E-5	14
NOP56	7.126E-5	13
ZNF516	8.39E-6	12
POLI	2.1447E-4	11
SYP	3.0982E-4	23
PGK1	1.05E-6	9
KDM1A	0.0	1
CCDC132	0.0	1
CDC27	0.0	1
KDM7A	9.0E-7	7
REV3L	3.7E-7	9
MBTPS2	0.0	6
ATP2B4	1.4088E-4	9
STAG3	2.1042E-4	9
SPEG	0.0	7
SF3B2	3.8E-7	9
RHBDD2	0.0	6
MYCBP2	1.9E-7	6
CRLF1	0.0	3
SYNRG	0.0	1
CACNG3	0.0	4
TMEM132A	0.0	1
RALA	0.0	7
AGK	0.0	3
GGCT	0.0	7
MARK4	0.0	6
PAFAH1B1	0.0	6

PTPN21	0.00650512	36
CACNA2D2	0.0	2
DNAJC11	0.0	5
PSMB1	0.0	1
CDKL5	0.0	1
MED24	0.0	1
HEATR5B	3.7E-7	7
SEC62	0.0	1
CSDE1	0.0	2
SEL1L	5.1E-7	4
BAZ1B	0.0	7
CLCN6	0.0	2
RFX2	1.3997E-4	20
CEP68	0.0	1
MAP4K3	0.0	7
EHD2	4.23E-6	10
KDM5D	0.0	1
UBR7	0.0	1
SLC7A14	0.0	1
RNF14	0.0	2
RGPD5	3.315E-5	19
MDH1	0.0	3
SLC30A9	0.0	1
COX15	0.0	2
PITHD1	1.1965E-4	9
CTSA	6.695E-5	3
TXNDC16	0.0	5
TFIP11	2.0645E-4	10
MYH7B	1.767E-5	14
RUFY3	0.0	3
ANK1	3.91E-6	9
IKZF2	0.0	7
VCAN	0.0	5
CDH10	2.83E-6	7
RAB27B	0.0	7
AP2S1	1.67E-6	7
CP	0.0	6
DKK3	7.27E-6	8
CDK17	5.19E-6	10
YBX3	0.0	1
ELMO2	1.1616E-4	8
PDE4A	0.0	7
IDI1	1.751E-5	9
ACSL4	1.67E-6	5
GNB5	0.0	5
DGCR2	1.0003E-4	8
ASNS	9.9E-7	8
OSBPL3	0.0	2
ATP6AP1	0.0	7
EIF4G3	0.0	1

USP33	0.0	8
OSBPL6	0.0	4
SMARCA2	4.04E-6	9
RSBN1	0.0	6
TNPO1	0.0	5
SMAP2	0.0	7
KIAA1467	0.0	7
RAB10	0.0	6
MAPRE3	0.0	2
SCAMP1	0.0	4
MTIF2	1.3303E-4	11
DDHD2	0.0	5
ATG16L1	0.0	7
IPO11	8.48E-6	9
NLK	1.67E-6	7
UIMC1	0.0	3
DNM1L	1.67E-6	5
EPB41L1	0.0	7
DOCK9	1.67E-6	5
DOCK3	0.0	7
TMEM230	7.27E-6	8
HNRNPC	1.21E-6	3
ARCN1	6.81E-6	4
CRTAC1	6.756E-5	11
POLRMT	1.21E-6	3
PACSIN2	7.27E-6	6
SAMM50	1.21E-6	3
KHNYN	1.025E-5	8
VTI1B	0.0	4
GMPR2	0.0	2
SAMHD1	0.0	4
KIF3B	0.0	5
TTI1	0.0	6
APMAP	4.04E-6	9
MCF2	7.27E-6	6
SMARCA1	1.751E-5	9
RTFDC1	0.0	4
GLRX2	0.0	1
RRAGD	0.0	6
PHF20	0.0	7
TOMM34	0.0	1
RTEL1-TNFRSF6B	0.0	8
AGPAT4	0.0	2
SLC39A9	2.4E-7	6
NUP160	0.0	1
RNF19A	0.0	1
SKIV2L2	1.9E-7	6
INPP4A	0.0	1
C12orf4	0.0	1

LMO3	0.0	2
MRPS10	0.0	1
MCUR1	0.0	1
LETMD1	0.0	3
MSMO1	0.0	2
MCF2L2	0.0	2
TAB2	2.4E-7	6
CCDC85A	0.0	4
ZFR	0.0	5
ATP11B	6.79E-6	9
RASGRF1	0.0	8
SLC2A3	0.0	1
CCAR1	0.0	4
LZTS1	0.0	1
MRPS35	0.0	1
CS	0.0	5
TAF2	0.0	1
TNPO3	0.0	1
OAT	0.0	1
KARS	0.0	1
TLE2	0.0	1
ASB1	0.0	2
NFYC	0.0	6
NGEF	0.0	4
ARFGEF1	0.0	9
PFKP	0.0	2
ATP2B3	6.695E-5	3
PDK3	0.0	1
COASY	0.0	1
TFE3	1.9E-7	6
TBC1D25	0.0	8
KIF2A	0.0	5
FUNDC1	0.0	3
TMEM260	1.9E-7	6
GBA2	0.0	5
AP3M2	0.0	1
ATP2B1	0.0	1
RPL31	1.9E-7	6
RPS6KA2	0.0	2
WBSCR22	0.0	2
PDCD2	1.9E-7	5
SLC6A15	0.0	1
TRHDE	0.0	1
CRMP1	1.9E-7	6
SDHA	0.0	4
SNCB	0.0	1
TSG101	6.758E-5	6
PTPLAD1	0.0	8
ENO1	0.0	2
ACTR6	0.0	5

SEMA3C	0.0	8
TIMM21	0.0	2
ADD2	7.15E-6	10
SEC31B	0.0	2
KIFAP3	0.0	1
RAP1GAP	0.0	2
DGKD	0.0	1
DNAJC10	0.0	4
JADE1	0.0	6
UBE2A	0.0	7
MAP2	7.15E-6	10
UBE2K	0.0	1
C12orf5	0.0	1
VDAC3	0.0	6
FDFT1	0.0	1
OPHN1	0.0	8
DDX1	0.0	4
OXCT1	0.0	1
STARD7	1.9E-7	4
COL16A1	0.0	4
TAF9	0.0	2
MECOM	0.0	1
CHMP5	0.0	3
ADD1	0.0	5
ATRN	1.3805E-4	9
RPL6	7.74E-6	11
ERP29	0.0	1
FUS	1.9E-7	6
GPATCH2L	0.0	1
CERS4	0.0	3
FCGBP	0.0	4
RBM27	0.0	4
DLD	0.0	2
ALKBH5	0.0	3
CD200	0.0	8
TOX4	0.0	1
DPYSL2	0.0	1
MAP3K1	0.0	1
ZNF184	1.06E-6	9
ABL1	0.0	1
RAB18	0.0	3
CEP170B	0.0	1
PALM	0.0	1
PICK1	0.0	2
RAB36	0.0	6
SBF1	0.0	1
DNAL4	0.0	1
TOM1	0.0	5
SYNGR1	1.9E-7	6
FAM118A	0.0	1

ACO2	0.0	8
DESI1	1.9E-7	6
CERK	0.0	3
DAAM1	7.0E-7	8
CHGA	0.0	5
DHRS7	0.0	3
PPM1A	0.0	1
KIAA0247	0.0	1
APEX1	0.0	1
BRMS1L	0.0	1
PYGB	0.0	7
ABHD12	0.0	1
PLCB4	0.0	1
MAP1LC3A	0.0	1
LPIN2	0.0	3
SMCHD1	0.0	1
ALG13	0.0	1
SARS	0.0	6
GABARAPL2	0.0	2
GPBP1	0.0	7
AFF4	0.0	5
UBE2D1	3.06E-6	8
MARK2	0.0	3
DLG1	0.0	8
ATXN7L3	0.0	3
KHSRP	0.0	3
CRKL	0.0	4
CHD8	0.0	4
C20orf24	0.0	5
TM9SF4	0.0	6
USP2	0.0	6
TUBG2	0.0	1
MYO16	2.85E-6	5
CUL7	0.0	5
PHKA2	0.0	1
ATP6V1H	0.0	4
KIAA0556	0.0	2
DTNBP1	0.0	1
HDAC9	8.181E-5	6
CLPTM1L	0.0	2
HEXB	0.0	1
KIAA2022	0.0	1
BCAR1	0.0	2
ARID4B	0.0	2
PTPRN	0.0	3
ATP9A	0.0	3
RC3H2	0.0	1
APPBP2	0.0	2
TM7SF3	0.0	5
BTBD1	0.0	6

SPEN	0.0	2
ROGDI	0.0	3
PSME4	0.0	2
MAST4	0.0	2
NUP133	0.0	3
CSNK2A2	0.0	3
RHOBTB1	0.0	3
ZZEF1	0.0	5
SEMA3A	0.0	1
TMEM131	1.1082E-4	8
ARHGEF1	0.0	4
REXO1	0.0	5
RIMS1	0.0	1
AP4E1	0.0	4
PCNP	0.0	5
ARG2	0.0	2
PLOD1	0.0	6
SEPHS1	0.0	3
DDX18	0.0	2
LZTS3	0.0	2
XRN2	0.0	2
ANAPC5	0.0	2
SLC23A2	0.0	1
NAT14	0.0	4
WDR7	0.0	4
FKBP5	0.0	6
CIRBP	0.0	4
SMARCB1	0.0	2
DDT	0.0	4
SEZ6L	0.0	3
DDX17	0.0	3
L3MBTL2	0.0	2
AP4S1	0.0	2
VCPKMT	0.0	3
TPD52L2	0.0	3
PSMA7	0.0	4
COL20A1	0.0	4
MANBAL	0.0	3
IDH3B	0.0	3
SNRNP40	0.0	3
SPHK2	0.0	1
PKN2	0.0	2
ADAM11	0.0	2
ACTN2	0.0	1
RBFOX1	0.0	1
MEF2C	0.0	2
C5orf22	0.0	3
SLC27A5	0.0	1
ACHE	0.0	1
NSFL1C	0.0	3

MAVS	0.0	4
PXN	0.0	1
GANAB	0.0	3
SNAP23	0.0	1
FBXL19	0.0	1
POLR2E	0.0	2
BCL2L13	0.0	1
CBX7	0.0	1
POLR3H	0.0	2
PSMD10	0.0	2
ZBTB11	0.0	1
IL4R	0.0	1
ERC1	0.0	1
GLG1	0.0	1
FH	0.0	2
BTAF1	0.0	1
JOSD1	0.0	1
ATP6V1D	0.0	1
VRK1	0.0	1
PGRMC1	0.0	1
SYT1	0.0	2
POLR1A	0.0	2
NUCKS1	0.0	2
C16orf80	0.0	1
PVALB	0.0	4
COL5A3	0.0	1
MAGI3	0.0	1
APOL4	0.0	2
CPNE6	0.0	1
FGFR1	0.0	1
NRD1	0.0	2
ALG9	0.0	1
MZF1	0.0	1
RBFOX2	0.0	1
ZC3H7B	0.0	1

Gene name	BetweennessCentrality	Degree
EPHA10	0.0	1
MGAT4C	0.0	1
WWP2	0.0	1
YY1AP1	0.0	1
PEA15	0.0	1
NF2	0.0	1
MDC1	0.0	1
KIAA1211L	0.0	3
GABRB1	0.0	3
PARVA	0.0	1
CXCL14	1.0019E-4	8
ZNF521	0.00690966	5
JUND	0.00284938	4
CTIF	0.0	3
MRAS	0.00877193	2
THRA	0.01746657	3
MEG8	0.0	2
CCNL1	0.0	3
IQGAP1	5.21E-6	6
SOCS4	8.9945E-4	4
COX20	5.0E-6	8
CRIP2	0.0	1
CUL9	0.0	1
SOWAHA	0.0	1
MAP3K5	0.0	2
NADK2	1.8063E-4	6
PPP1R9B	0.034624	2
KLF16	0.00672816	6
SLC25A23	0.02164694	7
MEIS3	0.0	4
XPC	0.0	3
CLIP1	5.72E-6	2
NLGN3	0.01411041	14
C14orf37	1.5489E-4	26
ZC3H7B	0.0	3
RPTOR	0.00196617	12
GTPBP1	0.0	4
TBC1D24	0.0	2
YPEL1	0.0	1
TFDP1	0.00809421	5
CTXN1	0.00112766	4
JPH4	0.00187388	3
FBXO41	0.06160946	18
CBX7	5.7885E-4	10
RN7SL4P	0.01877667	4
ANP32E	0.00760789	6
MAPK4	2.2142E-4	4
GCSH	0.0	2
FAM184A	0.0212699	6
HADHA	0.00150989	4
IVD	0.01978902	12
TDRD3	0.0	1
TMEM219	0.0	1
CCND3	0.02616122	3
KCNN2	0.00877193	2
MAGI2	6.44E-6	7
SLITRK2	0.00930428	6
CADM1	8.4598E-4	8
EPHB1	1.456E-4	6
MEIS2	8.2E-6	5
BBS2	8.2E-6	5
MARC2	7.058E-5	9
CHRDL1	0.0	3
PGM1	0.0180841	18
HDAC5	7.67E-6	5
SNX8	2.03E-6	3
ARHGAP10	0.0	3
PGGT1B	0.0	2
TMBIM6	0.09165216	31
HERPUD2	7.8312E-4	14
CPE	0.00833636	16
ACOX1	0.00105274	26
WASF3	0.00988355	26
CSPG5	0.00258724	28
LAPTM4A	0.0	4
KDM3B	0.01999232	25
BRINP2	0.00718385	52
KLHL26	3.163E-5	17
PEX10	0.00503839	27
FSIP1	0.00337346	16
SPAG16	0.0	2
PAK6	2.2776E-4	25
MRPS14	1.2223E-4	21
KCNIP2	0.0	4
HNRNPUL1	0.0	4
TRA2A	3.86E-6	3
PDXK	0.0	4
FMN2	0.01607117	7
PYCRL	0.0	3
RAB36	0.0044059	9
CRAT	0.04154943	15
TMEM259	1.1926E-4	5
HCFC1	0.03097172	9
JPH3	0.00407582	8
NEURL1	2.6434E-4	7
LZTS1	0.0	4
GATS	0.0	2
DCAKD	7.805E-5	6
SEC61A1	2.1228E-4	4

NPTXR	0.01890967	45
TLE1	9.794E-5	30
NELFB	0.0045611	24
EIF3F	0.0	1
ST6GALNAC6	1.2905E-4	15
PFKL	0.0	2
STK11	0.00976168	40
APH1A	3.92E-6	14
PPM1F	0.00556671	30
TFE3	0.01137781	28
MED29	0.02131575	15
SLC6A1	5.8522E-4	14
SNX1	1.47E-6	11
SPG7	2.17E-6	12
C11orf95	0.00102062	42
DYNC2H1	1.0618E-4	24
ST6GAL2	0.00413748	41
SEPT11	0.03833608	63
HS6ST1	0.0	14
LINC00094	0.00120466	57
RP11-82L18.4	6.9876E-4	50
SCAF8	0.02907944	61
C15orf59	0.00447618	32
RP11-514P8.6	5.742E-5	15
GIGYF2	6.6382E-4	61
SHISA4	0.00247505	47
R3HDM4	0.00450116	62
C1orf122	0.01196448	79
IPP	0.00803934	48
MAPK11	0.00612578	36
SEPT5	0.02134231	111
CBX6	0.03620817	88
KCNIP1	0.0	6
SETD2	5.03E-6	29
RBM15B	0.02298066	100
LINC00086	0.01351673	92
RIMS2	7.71E-6	26
ZBTB4	0.00485158	64
CCDC106	0.01430967	60
GOLGB1	0.00482071	52
ADRBK1	0.00117822	47
CES2	0.00126766	13
GNB2	4.91E-5	25
ORMDL3	8.82E-5	38
SYNPO	0.01212332	66
CTNND2	0.01280971	74
NYAP1	0.00209014	19
FASTK	2.999E-5	23

HIPK1	0.0257471	96
PMVK	0.00249643	39
AGL	3.6264E-4	19
EMC10	0.005544	63
FBXW5	1.308E-5	19
FAM213B	0.00969789	66
PRKCA	0.00554267	32
TMCO3	0.00513277	61
SHC3	0.00851982	57
SNRNP200	0.01502858	108
PTPRF	0.0191104	54
RERE	2.528E-5	35
AKT1	0.0	21
LARP1B	3.4725E-4	42
APH1B	0.02716877	73
SLTM	8.08E-6	30
ARRB1	4.5083E-4	41
APC	9.4379E-4	39
MYBBP1A	0.01205141	55
SYNE1	0.00738708	21
KIF1A	0.0012777	38
AKAP9	0.0	6
SMARCA4	0.00425928	13
MMP24	2.7295E-4	44
PACSIN1	0.01561015	85
OBSL1	2.6448E-4	34
ZC3H13	7.0677E-4	40
CLCC1	1.8314E-4	38
MAPK8IP1	0.00528651	84
DARS	0.0216334	69
SRF	3.865E-5	34
NCS1	9.5741E-4	28
AK1	0.01593364	108
KXD1	0.0	13
SGTA	0.01749445	101
WDR13	0.00767213	93
ANKRD12	7.736E-5	37
SNPH	0.00652234	18
RIMS4	0.0048375	58
CPNE6	3.66E-6	16
ITPK1	5.2324E-4	55
PITPNM3	0.02871149	58
RAPGEF4	7.67E-6	7
PCM1	0.00394032	30
CAMK2A	0.01642707	63
PDZD4	0.01137765	78
ZNF275	2.669E-5	31
TPR	0.02578239	71
RETSAT	5.61E-5	14
LARS2	2.248E-5	26

SEMA6B	0.0051309	46
LZTS3	0.02715887	33
DBNDD1	6.8E-6	5
STARD13-AS	5.2254E-4	29
FXYD7	0.04373953	134
FAM19A5	0.00160386	53
EHMT2	0.00603833	62
PDE7B	2.42E-6	16
KLF13	0.00124366	69
ROBO1	0.02410722	128
SIN3A	0.01903934	116
HSD11B1L	1.6031E-4	45
KIAA1429	0.0111327	106
BAP1	0.0396547	127
RPS6KA4	6.9142E-4	54
MUM1	0.00611314	80
NACC1	0.01072852	97
SH3GLB2	0.00559721	93
SLC25A26	0.00374264	70
FEM1A	0.01185686	98
WDR89	0.00933403	102
SYTL2	1.4872E-4	42
BAI3	0.0155098	94
ATP5S	9.135E-5	31
SLC25A16	0.00976257	79
MAPKAP1	0.00669577	101
B4GALT2	5.8287E-4	50
RLF	0.00637463	90
NCBP2	0.02144066	122
NECAB2	3.9146E-4	41
PCIF1	0.02797204	113
SMARCB1	0.00843432	107
FBXL19	0.00627988	73
SEPHS1	0.01016883	106
DGCR2	0.015682	115
DPF1	0.00425826	86
THSD7A	0.01809894	112
BAD	1.1029E-4	34

Gene name	BetweennessCentrality	Degree
CD38	6.05E-5	9
TPR	0.03210921	151
FUBP3	0.00149359	16
RAB5C	0.0109917	73
CCDC88A	0.00618185	35
CYB5R1	0.09987836	199
PAQR6	0.00287029	25
RMDN1	0.00200798	25
IPP	0.04223124	108
FAM229B	0.0010412	29
FAM214B	3.0087E-4	17
MED29	0.01113747	36
RNF31	0.00193693	62
MARCH2	8.5325E-4	44
ZKSCAN1	0.02889723	134
SUMF2	0.06466922	189
DIRC2	6.8827E-4	43
TCTA	8.0335E-4	28
GPR107	0.03555079	142
AGPAT5	0.0432596	94
IDH2	0.00429579	37
NA	0.04270091	167
RP11-617F23.1	0.00243147	32
RNF216	2.26E-6	14
HMGB3	0.00584314	83
TRMU	5.0809E-4	36
SSR3	0.02943845	131
ABCD4	0.01209497	111
GIPC1	0.05120506	176
HSD17B12	0.02611942	152
RNF20	0.08244883	193
FGF17	0.01124752	105
CTSF	0.04340171	163
HECW1-IT1	0.08715323	177
RP11-334C17.5	0.00420764	75
CLK1	1.979E-5	16
PIGS	2.0346E-4	28
C20orf194	0.00350117	74
GORASP1	0.00148992	57
ACRC	0.00275498	74
SPAG5-AS1	0.01684795	118
CD44	0.0	7
CLNS1A	0.01214625	110
MFSD8	0.02975281	147
TMEM9B	0.03874169	160
RETSAT	6.448E-5	23
ACAA1	3.9495E-4	21
MAOB	0.0	8
TFRC	0.0	5
ANO8	3.871E-5	17
PAPOLA	1.74E-5	28
LMF2	2.09E-6	22
TXN2	6.029E-5	29
ZNF629	2.665E-5	22
SCG3	1.502E-5	23
PYCRL	3.521E-5	24
PTOV1	1.114E-4	28
ETFB	0.00286679	78
SNX8	8.17E-6	24
AK1	9.48E-6	17
ALDOC	0.00135121	32
ATG2A	1.2656E-4	25
DARS	3.8031E-4	41
WDR35	0.00623601	103
ACOT2	0.0	15
DNAJC15	1.59E-6	19
FKBP9	3.052E-5	25
PFKFB2	1.5904E-4	33
PCID2	0.01846826	50
SLC5A6	1.267E-5	21
PYCR2	0.0051576	84
KANSL1L	3.46E-6	16
GBAS	7.074E-4	45
TAF1	3.57E-6	19
ALAD	0.00330413	79
IMPACT	0.00948601	82
VOPP1	2.2854E-4	29
SLC6A1	4.925E-5	26
LRP8	1.23E-6	14
MX1	5.761E-5	30
CCDC117	3.88E-5	22
SHISA5	0.0	11
SGK494	2.4493E-4	38
TADA3	2.5986E-4	41
THAP2	0.0	6
RCC1	4.435E-5	31
HIST1H2BC	0.01582267	113
C8orf33	1.37E-5	19
ZBTB40	1.335E-5	26
P4HB	6.4252E-4	48
NAT8L	9.22E-6	21
HNRNPU-AS1	1.0805E-4	31
AMZ2	1.881E-5	23
COX20	4.893E-5	28
RBM14	3.1904E-4	37
GATS	1.2504E-4	32

MRPS6	3.507E-5	19
RNU6-6P	2.81E-6	10
SLC5A3	3.12E-6	16
RANBP3	3.28E-6	13
GNS	2.1155E-4	37
MIR137HG	0.00558295	50
TNC	0.0	9
ASF1A	0.0	16
DCUN1D1	0.0	2
MAP4	6.414E-5	28
MXD1	1.6816E-4	28
SRBD1	9.805E-5	23
JMJD6	7.226E-5	19
NDST1	3.9172E-4	15
SREBF1	3.583E-4	15
SPAG5	8.344E-5	29
PCM1	1.302E-5	15
TDRD3	1.0816E-4	22
TAB1	5.926E-5	26
COTL1	0.0	4
MYH14	3.475E-4	18
KDELR1	1.771E-5	12
PDE4C	5.307E-5	24
DFNA5	4.681E-5	32
ACBD5	1.47E-4	31
RAI1	8.4E-7	18
KAT2A	1.313E-5	12
CHD4	1.0724E-4	30
BAG2	2.173E-5	12
ACOT13	4.195E-5	23
EPM2A	1.4212E-4	37
CSPG5	1.223E-5	26
MRPL19	0.0032862	73
KLHL29	4.58E-6	22
MRPS14	0.0	11
FAM213A	1.47E-5	16
RAP2C	8.981E-5	33
HIVEP3	0.02312267	43
UCK1	2.171E-5	11
GSTM3	0.0	7
DHX34	8.738E-5	30
KCTD3	9.601E-5	15
MYO7A	1.93E-6	19
APH1B	0.00299837	35
SLC7A1	2.98E-6	21
ETFA	2.662E-5	26
SEMA6C	3.576E-5	22
SPAG16	2.339E-5	25
EIF5B	0.0	12
TK2	3.635E-5	18
FTH1	2.44E-6	22
C11orf84	1.696E-5	17
SIN3A	4.647E-5	20
BANP	1.8077E-4	32
CLK3	6.54E-6	23
SKA2	9.729E-5	22
NRBP2	1.899E-5	21
NELFB	0.0	12
TTC30A	2.2335E-4	35
SLC9A8	5.54E-5	22
C1orf122	3.735E-5	27
SHISA4	6.3111E-4	30
BCYRN1	0.0	12
STRC	0.01258639	43
BCKDHA	2.2006E-4	38
LINC00599	1.4671E-4	28
RP11-282K24.3	6.97E-6	21
RP11-192H23.4	1.07E-5	25
RP11-252A24.7	0.0	13
SUZ12P	8.89E-6	21
SNX29	0.0	4
CRAT	0.0	5
IVD	0.0	5
WASF3	0.0	1
ZNF710	7.47E-6	6
BAALC	0.0	6
AF127936.7	7.8E-6	11
TBC1D1	5.38E-6	19
CTNNA2	0.0	3
TPPP3	2.0714E-4	31
TRIB2	0.0	2
DAZAP1	0.0	7
SPECC1L	0.0	18
CRYBB2P1	0.0	15
MON1B	2.27E-6	12
ECH1	3.75E-6	16
SPATA6L	5.18E-6	19
NCBP2	4.19E-6	18
H3F3B	0.0	9
APLNR	0.0	9
SH3GL1	0.0	17
KCNE4	7.31E-6	19
EMC10	1.969E-5	18
TRIM17	1.389E-5	23
SETD5	0.0	9
ZNF154	0.0	12
LINC00925	0.0	5

MARK3	0.0	5
TP53INP2	2.7742E-4	18
LAMTOR1	6.9948E-4	21
HDAC11	3.3234E-4	20
DLGAP4	0.0	3
ZC3H7A	1.801E-5	17
COQ9	0.0	3
GTPBP1	7.68E-6	19
PCYT1B	0.0	12
CHPF	0.0	13
SPARCL1	0.0	12
SDHAP1	1.75E-6	11
TMEM222	1.954E-5	22
PITPNM2	0.0	3
MYO15A	0.0	3
HIF1A	0.0	1
CDC25B	0.0	2
HSF4	6.1E-7	15
ZBTB16	0.0	14
MAST1	0.0	1
DNAJC2	0.0	7
SNX19	0.0	9
THEM6	0.0	4
IPO8	0.0	8
AKT1	0.0	1
MTURN	0.0	10
MEGF9	0.0	3
NLGN3	1.464E-5	5
RPL28	0.0	1
RBP1	6.114E-5	10
ADAM15	7.411E-5	8
ALDH1A1	1.294E-5	4
PLXNA1	0.0	3
PASK	8.661E-5	16
DGCR8	0.0	4
USP21	0.0	6
TRA2A	7.99E-6	13
LINC00632	0.00755098	24
MBOAT2	1.617E-5	4
BEST1	0.0	3
SLC35A4	0.0	4
ZNF808	2.884E-5	4
STAG3L5P-PVRIG2P-PILRB	0.0	6
KIF21B	0.0	2
FAM104A	0.0	8
RP11-1212A22.1	0.0	7
CSMD2	9.7911E-4	10

B4GALNT3	0.0	3
C2orf49	0.0	7
GPR153	0.0	5
C21orf2	0.0	2
MGAT5B	0.0	7
FAM153B	0.0	14
CSPG4P11	0.0	5
FAM210B	0.0	1
HSD17B7	0.0	1
SNX32	1.134E-5	4
MTUS2	0.0	2
RP11-1407O15.2	0.0	2
MIAT	1.313E-5	7
EYA3	0.0	4
PIGP	0.0	2
PNPLA7	0.0	3
LMBR1L	0.0	1
TMEM108	0.0	2
HS2ST1	0.0	3
DDAH1	0.0	1
GAN	0.0	1
TNFAIP8L1	0.0	1

A.3 R-files

- Datastep.R
- StepByStepWGCNA.R
- Analysis.R
- Cytoscape.R

```
1
2  library("tidyverse") #package for dplyr
3  library("WGCNA")
4
5  Cohorts <- c("PW", "NBB", "PA")
6
7  # Genenames
8  # geneNames <- readRDS("../Data/EnsDb.Hsapiens.v75.Rds")
9
10 # Sample info
11 SampleInfo <- read.csv(paste0("../Data/sample_list_3cohorts_final.csv"), header=TRUE, sep = "\t"
       )
12
13 # Count matrix
14 countMatrix <- read.csv(paste0("../Data/countMatrix.genes"), header=TRUE, sep = "\t")
15
16 # First, we want rows to correspond to samples and columns to genes, we
17 # transpose the matrix
18 counts_t <- t(as.matrix(countMatrix[,-1]))
19 colnames(counts_t) <- countMatrix$genes
20
21 #
22 # Function to extract the counts for a specific cohort and transpose the count
23 # matrix
24 #
25 countsCohort <- function(cmat, mdat, cohort) {
26   tmpMat <- cmat %>% select_if(names(.) %in% mdat[mdat$origin==cohort,"sample_id"]) %>%
27     as.matrix %>% round %>% t
28   colnames(tmpMat) <- cmat$genes
29   return(tmpMat)
30
31 }
32
33 # EXAMPLE: extract counts from PW cohort
34 count_PW <- countsCohort(countMatrix, SampleInfo, "PW") #%>% .[1:10,1:10]
35 count_PW_example <- countsCohort(countMatrix, SampleInfo, "PW") %>% .[1:10,1:1000] %>% as.matrix
36
37
38 # CLEANING DATASET
39 #
40 # Find genes with too many missing values and remove them from the count
41 # matrix. We do this for every cohort in an "lapply" loop to have a list with
42 # the results for each cohort
43 #
44 counts <- lapply(Cohorts, function(cohort){
45   tmpCounts <- countsCohort(countMatrix, SampleInfo, cohort)
46   gsg <- goodSamplesGenes(tmpCounts, verbose=3)
47   tmpCounts[gsg$goodSamples, gsg$goodGenes]
48 })
49 names(counts) <- Cohorts
50
51 #
52 # CLUSTERING SAMPLES
53 #
54 # To remove outlier samples, first we cluster them
55 #
```

```
# 1) PW cohort
#
cohort <- "PW"
sampleTree <- hclust(dist(counts[[cohort]]), method="average")
plot(sampleTree, main=paste0(cohort, " cohort"), sub="", xlab="")
#
# By looking at the plot, there is at least one oulier, the sample SL283597. We
# chose a threshold of 3000000 to cut the sample out, and plot the line on the
# dendrogram to be sure.
#
cutH <- 3000000
abline(h=cutH, col="red")
clust <- cutreeStatic(sampleTree, cutHeight=cutH, minSize=10)

# There should be 2 clusters, we want to keep the one with 27 samples, of
# course, and discard the sample that we cut out
#
table(clust)

# So, the cluster we want to keep is the [1], let's get the sample ids
#
keepSamples <- rownames(counts[[cohort]])[clust==1]

# Now we can update the counts matrix and the SampleInfo by removing the
# outlier (we are going to create a list of sample_info, with one element for
# each cohort)
#
counts[[cohort]] <- counts[[cohort]][rownames(counts[[cohort]]) %in% keepSamples,]
metadata <- list()
metadata[[cohort]] <- SampleInfo[SampleInfo$sample_id %in% keepSamples,]


# SELECT MOST VARYING GENES
#
# This function returns the top X most varying genes
#
topMad <- function(cmat, top=10000) {
  apply(cmat, 2, mad) %>% sort(decreasing=TRUE) %>% head(n=top) %>% names
}

# We select the top 10000 most varying genes for each of the three cohorts
#
topN <- 10000
counts_top <- lapply(Cohorts, function(cohort){
  top_ids <- topMad(counts[[cohort]], topN)
  counts[[cohort]][,colnames(counts[[cohort]]) %in% top_ids]
})
names(counts_top) <- Cohorts

counts_top_PW <- counts_top[["PW"]]


# SUBSETS BASED ON CONDITION
#
#Function
#
countsCondition <- function(cmat, mdat, cond) {
  tmpMat <- cmat[rownames(cmat) %in% mdat[mdat$condition==cond,"sample_id"],] %>%
    as.matrix %>% round
  return(tmpMat)

}

counts_PW_control <- countsCondition(counts_top_PW, SampleInfo, "Control")
counts_PW_case <- countsCondition(counts_top_PW, SampleInfo, "Case")


# SOFT THRESHOLD
#
```

```
#Function
#
softThreshold <- function(countMatrix_cohort){
  #Settings
  options(stringsAsFactors = FALSE)

  #Choose a set of soft -threshold powers
  powers = c(c(1:10), seq(from = 12, to = 30, by=2))

  #Calling the network topology analysis funtion
  sft = pickSoftThreshold(countMatrix_cohort, powerVector = powers, verbose = 5)

  #Plot the results of the analysis
  sizeGrWindow(9,5)
  par(mfrow = c(1,2))
  cex1 = 0.8

  #Plot, scale-free topology fit index
  plot(sft$fitIndices[,1], -sign(sft$fitIndices[,3])*sft$fitIndices[,2], xlab = "Soft Threshold
       (power)", ylab = "Scale Free Topology Model Fit, signed R^2"
       , type="n", main = paste("Scale independence"))
  text(sft$fitIndices[,1], -sign(sft$fitIndices[,3])*sft$fitIndices[,2], labels = powers, cex =
       cex1, col="red")

  abline(h= cex1, col= "red")

}

softThreshold(counts_PW_case)
softThreshold(counts_PW_control)

softThreshold(count_PW_example)
```

Rfiles/Datastep.R

```
library(WGCNA)

# CONTROL
adjacency= adjacency(counts_PW_control, power = 24)

#Turning adjacency into topological overlap
TOM_control = TOMsimilarity(adjacency)
dissTOM_control = 1-TOM_control

#Clustering using TOM
#Call the hierarchical clustering function
geneTree_PW_control = hclust(as.dist(dissTOM_control), method = "average")

#plot the tree
sizeGrWindow(12,9)
plot(geneTree_PW_control, xlab = "", sub = "", main = "Gene clustering on TOM-based
     dissimilarity", labels = FALSE, hang = 0.4)

#Branch cutting and modules
#Min module size
minModuleSize = 30

#Module identification using dynamic tree cut:
dynamicMods_control = cutreeDynamic(dendro = geneTree_PW_control, distM = dissTOM_control,
     deepSplit = 2, pamRespectsDendro = FALSE, minClusterSize = minModuleSize)
table(dynamicMods_control)

#Convert numeric labels into colors
dynamicColors_control = labels2colors(dynamicMods_control)
table(dynamicColors_control)

#Plot the dendogram with colors underneath
```

```
plotDendroAndColors(geneTree_PW_control, dynamicColors_control, "Dynamic Tree Cut", dendroLabels = FALSE, hang = 0.03, addGuide = TRUE, guideHang = 0.05,
                    main = "Gene dendrogram and module colors")

#Module eigengene- merging of modules whose expression profiles are very similar
#Calculate eigengenes
MEList_control = moduleEigengenes(counts_PW_control, colors = dynamicColors_control)
MEs_control = MEList_control$eigengenes

#Calculate dissimilarity of module eigengenes
MEDiss_control = 1-cor(MEs_control)

#Cluster module eigengenes
METree_control = hclust(as.dist(MEDiss_control), method = "average")

#Plot the result
sizeGrWindow(7,6)
plot(METree_control, main = "Clustering of module eigengenes, Control", xlab = "", sub = "")

#Choosing cut,
MEDissThres = 0.1

#plot the cut line
abline(h= MEDissThres, col = "red")

#Call an automatic merge function
merge_control = mergeCloseModules(counts_PW_control, dynamicColors_control, cutHeight = MEDissThres, verbose = 3)

#The merged module colors
mergedCOlors_control = merge_control$colors
table(mergedCOlors_control)

#Eigengenes of the new merged modules
mergedMEs_control = merge_control$newMEs

#plot the gene dendrogram again
sizeGrWindow(12,9)
plotDendroAndColors(geneTree_PW_control, cbind(dynamicColors_control, mergedCOlors_control), c("Dynamic Tree Cut", "Merged dynamic"),
                    dendroLabels = FALSE, hang = 0.3, addGuide = TRUE, guideHang = 0.05)

#Topological overlap, using dissTOM
#Transform dissTOM with a power to make moderaely strong connections more visible in the heatmap
plotTom = dissTOM_control^7

#Set diagonal to NA for a nicer plot
diag(plotTom) = NA

#Call the plot function
sizeGrWindow(9,9)
TOMplot(TOM_control, geneTree_PW_control, mergedCOlors_control, main = "Network heatmap plot, all genes")


################################################################################
#CASE
adjacency= adjacency(counts_PW_case, power = 5)

#Turning adjacency into topological overlap
TOM_case = TOMsimilarity(adjacency)
dissTOM_case = 1-TOM_case

#Clustering using TOM
#Call the hierarchical clustering function
geneTree_PW_case = hclust(as.dist(dissTOM_case), method = "average")

#plot the tree
sizeGrWindow(12,9)
```

```
plot(geneTree_PW_case, xlab = "", sub = "", main = "Gene clustering on TOM-based dissimilarity", labels = FALSE, hang = 0.4)

#Branch cutting and modules
#Min module size
minModuleSize = 30

#Module identification using dynamic tree cut:
dynamicMods_case = cutreeDynamic(dendro = geneTree_PW_case, distM = dissTOM_case, deepSplit = 2, pamRespectsDendro = FALSE, minClusterSize = minModuleSize)
table(dynamicMods_case)

#Convert numeric labels into colors
dynamicColors_case = labels2colors(dynamicMods_case)
table(dynamicColors_case)

#Plot the dendogram with colors underneath
plotDendroAndColors(geneTree_PW_case, dynamicColors_case, "Dynamic Tree Cut", dendroLabels = FALSE, hang = 0.03, addGuide = TRUE, guideHang = 0.05,
                    main = "Gene dendrogram and module colors")

#Module eigengene- merging of modules whose expression profiles are very similar
#Calculate eigengenes
MEList_case = moduleEigengenes(counts_PW_case, colors = dynamicColors_case)
MEs_case = MEList_case$eigengenes

#Calculate dissimilarity of module eigengenes
MEDiss_case = 1-cor(MEs_case)

#Cluster module eigengenes
METree_case = hclust(as.dist(MEDiss_case), method = "average")

#Plot the result
sizeGrWindow(7,6)
plot(METree_case, main = "Clustering of module eigengenes, Case", xlab = "", sub = "")

#Choosing cut,
MEDissThres = 0.2

#plot the cut line
abline(h= MEDissThres, col = "red")

#Call an automatic merge function
merge_case = mergeCloseModules(counts_PW_case, dynamicColors_case, cutHeight = MEDissThres, verbose = 3)

#The merged module colors
mergedColors_case = merge_case$colors
table(mergedColors_case)

#Eigengenes of the new merged modules
mergedMEs_case = merge_case$newMEs

#plot the gene dendrogram again
sizeGrWindow(12,9)
plotDendroAndColors(geneTree_PW_case, cbind(dynamicColors_case, mergedColors_case), c("Dynamic Tree Cut", "Merged dynamic"),
                    dendroLabels = FALSE, hang = 0.3, addGuide = TRUE, guideHang = 0.05)

#Topological overlap, using dissTOM
#Transform dissTOM with a power to make moderaely strong connections more visible in the heatmap
plotTom = dissTOM_case^7

#Set diagonal to NA for a nicer plot
diag(plotTom) = NA

#Call the plot function
sizeGrWindow(9,9)
TOMplot(TOM_case, geneTree_PW_case, mergedColors_case, main = "Network heatmap plot, all genes")
```

```
161
162 ##############################################################
163 # Example
164 adjacency= adjacency(count_PW_example , power = 14)
165
166 #Turning adjacency into topological overlap
167 TOM_example = TOMsimilarity(adjacency)
168 dissTOM_example = 1-TOM_example
169
170 #Clustering using TOM
171 #Call the hierarchical clustering function
172 geneTree_PW_example = hclust(as.dist(dissTOM_example), method = "average")
173
174 #plot the tree
175 sizeGrWindow(12,9)
176 plot(geneTree_PW_example, xlab = "", sub = "", main = "Gene clustering on TOM-based
        dissimilarity", labels = FALSE, hang = 0.4)
177
178 #Branch cutting and modules
179 #Min module size
180 minModuleSize = 30
181
182 #Module identification using dynamic tree cut:
183 dynamicMods_example = cutreeDynamic(dendro = geneTree_PW_example, distM = dissTOM_example,
        deepSplit = 1, pamRespectsDendro = FALSE, minClusterSize = minModuleSize)
184 table(dynamicMods_example)
185
186 #Convert numeric labels into colors
187 dynamicColors_example = labels2colors(dynamicMods_example)
188 table(dynamicColors_example)
189
190 #Plot the dendogram with colors underneath
191 #sizeGrWindow(8,6)
192 plotDendroAndColors(geneTree_PW_example, dynamicColors_example, "Dynamic Tree Cut", dendroLabels
        = FALSE, hang = 0.03, addGuide = TRUE, guideHang = 0.05,
193                     main = "Gene dendrogram and module colors")
194
195 #Module eigengene- merging of modules whose expression profiles are very similar
196 #Calculate eigengenes
197 MEList_example = moduleEigengenes(count_PW_example, colors = dynamicColors_example)
198 MEs_example = MEList_example$eigengenes
199
200 #Calculate dissimilarity of module eigengenes
201 MEDiss_example = 1-cor(MEs_example)
202
203 #Cluster module eigengenes
204 METree_example = hclust(as.dist(MEDiss_example), method = "average")
205
206 #Plot the result
207 sizeGrWindow(7,6)
208 plot(METree_example, main = "Clustering of module eigengenes", xlab = "", sub = "")
209
210 #Choosing cut,
211 MEDissThres = 1
212
213 #plot the cut line
214 abline(h= MEDissThres, col = "red")
215
216 #Call an automatic merge function
217 merge_example = mergeCloseModules(count_PW_example, dynamicColors_example, cutHeight =
        MEDissThres, verbose = 3)
218
219 #The merged module colors
220 mergedCOlors_example = merge_example$colors
221 table(mergedCOlors_example)
222
223 #Eigengenes of the new merged modules
224 mergedMEs_example = merge_example$newMEs
225
```

```
#plot the gene dendrogram again
sizeGrWindow(12,9)
plotDendroAndColors(geneTree_PW_example, cbind(dynamicColors_example, mergedCOlors_example), c("
    Dynamic Tree Cut", "Merged dynamic"),
                    dendroLabels = FALSE, hang = 0.3, addGuide = TRUE, guideHang = 0.05)

#Topological overlap, using dissTOM
#Transform dissTOM with a power to make moderaely strong connections more visible in the heatmap
plotTom = dissTOM_example^7

#Set diagonal to NA for a nicer plot
diag(plotTom) = NA

#Call the plot function
sizeGrWindow(9,9)
TOMplot(TOM_example, geneTree_PW_example, mergedCOlors_example, main = "Network heatmap plot,
    all genes")
```

Rfiles/StepByStepWGCNA.R

```
##############CORRESPONDENCE MATRIX: Healthy VS Control
#Preparing
caseMEs = orderMEs(mergedMEs_case, greyName = "MEO")
controlMEs=orderMEs(mergedMEs_control, greyName = "MEO")

#Isolating the module labels in the order they appear in ordered ME, already in colors
caseModules = substring(names(caseMEs),3)
controlModules = substring(names(controlMEs), 3)

#Initialize tables of p-values and of the corresponding counts
pTable = matrix(0, nrow = length(caseModules), ncol = length(controlModules))
CountTbl = matrix(0, nrow = length(caseModules), ncol = length(controlModules))

# Execute all pairwaise comparisons
for (camod in 1:length(caseModules))
  for (comod in 1:length(controlModules))
  {
    caseMembers = (mergedColors_case == caseModules[camod])
    controlMembers = (mergedCOlors_control == controlModules[comod])
    pTable[camod, comod] = -log10(fisher.test(caseMembers, controlMembers, alternative = "
        greater")$p.value)
    CountTbl[camod, comod] = sum(mergedColors_case == caseModules[camod] & mergedCOlors_control
        ==controlModules[comod])
  }

##To add colors

# Truncate p values smaller than 10^{-50} to 10^{-50}
pTable[is.infinite(pTable)] = 1.3*max(pTable[is.finite(pTable)])
pTable[pTable>50 ] = 50

# Marginal counts (really module sizes)
caseModTotals = apply(CountTbl, 1, sum)
controlModTotals = apply(CountTbl, 2, sum)

# Actual plotting
sizeGrWindow(12,20)
par(mfrow=c(1,1))
par(cex = 1.0)
par(mar=c(8, 10.4, 2.7, 1)+0.3)

# Use function labeledHeatmap to produce the color-coded table with all the trimmings
labeledHeatmap(Matrix = pTable,
               xLabels = paste(" ", controlModules),
               yLabels = paste(" ", caseModules),
               colorLabels = TRUE,
               xSymbols = paste("Control ", controlModules, ": ", controlModTotals, sep=""),
               ySymbols = paste("Case ", caseModules, ": ", caseModTotals, sep=""),
```

```
47                textMatrix = CountTbl,
48                colors = greenWhiteRed(100)[50:100],
49                main = "Correspondence of case vs control modules",
50                cex.text = 1.0, cex.lab = 1.0, setStdMargins = FALSE)
51
52
53 ######MODULE EIGENGENE HEATMAP: Visualizing network of eigengenes
54 #Plot relationships among the eigengenes
55 # Plot the heatmap matrix
56 par(cex = 1.0)
57 plotEigengeneNetworks(mergedMEs_case, "Eigengene adjacency heatmap, Case", marHeatmap = c
       (3,4,2,2),
58                       plotDendrograms = FALSE, xLabelsAngle = 90)
59
60 # Plot the heatmap matrix (note: this plot will overwrite the dendrogram plot)
61 par(cex = 1.0)
62 plotEigengeneNetworks(mergedMEs_control, "Eigengene adjacency heatmap, Control", marHeatmap = c
       (3,4,2,2),
63                       plotDendrograms = FALSE, xLabelsAngle = 90)
64
65 #
66 # MODULE PRESERVATION
67 #
68 #Multidata
69 countMatrix_2dim <- list(Control = list(data = counts_PW_control), Case = list(data = counts_PW_
       case))
70
71 #Colors
72 mergedColors_2dim <- list(Control = mergedCOlors_control, Case = mergedColors_case)
73
74 #ModulePreservation
75 mp <- modulePreservation(multiData = countMatrix_2dim, multiColor =  mergedColors_2dim,
76                          referenceNetworks = 1, nPermutations = 200, randomSeed = 1, quickCor =
                               0, verbose = 3)
77 #Analysis
78 #Isolating the observed statistics and their Z scores
79 ref = 1
80 test = 2
81
82 statsObs = cbind(mp$quality$observed[[ref]][[test]][, -1], mp$preservation$observed[[ref]][[test
       ]][, -1])
83 statsZ = cbind(mp$quality$Z[[ref]][[test]][, -1], mp$preservation$Z[[ref]][[test]][, -1])
84
85 # Module labels and module sizes are also contained in the results
86 modColors = rownames(mp$preservation$observed[[ref]][[test]])
87 moduleSizes = mp$preservation$Z[[ref]][[test]][, 1];
88
89 # leave grey modules out
90 plotMods = !(modColors %in% c("grey"));
91
92 # Text labels for points
93 text = modColors[plotMods];
94
95 # Auxiliary convenience variable
96 plotData = cbind(mp$preservation$observed[[ref]][[test]][, 2], mp$preservation$Z[[ref]][[test
       ]][, 2])
97
98 # Main titles for the plot
99 mains = c("Preservation Median rank", "Preservation Zsummary");
100
101 # Start the plot
102 sizeGrWindow(10, 5);
103 par(mfrow = c(1,2))
104 par(mar = c(4.5,4.5,2.5,1))
105 for (p in 1:2)
106 {
107   min = min(plotData[, p], na.rm = TRUE);
108   max = max(plotData[, p], na.rm = TRUE);
109   # Adjust ploting ranges appropriately
```

```
  if (p==2)
  {
    if (min > -max/10) min = -max/10
    ylim = c(min - 0.1 * (max-min), max + 0.1 * (max-min))
  } else
    ylim = c(max + 0.1 * (max-min), min - 0.1 * (max-min))
  plot(moduleSizes[plotMods], plotData[plotMods, p], col = 1, bg = modColors[plotMods], pch =
      21,
       main = mains[p],
       cex = 2.4,
       ylab = mains[p], xlab = "Module size", log = "x",
       ylim = ylim,
       xlim = c(10, 2000), cex.lab = 1.2, cex.axis = 1.2, cex.main =1.4)
  labelPoints(moduleSizes[plotMods], plotData[plotMods, p], text, cex = 1, offs = 0.08);
  # For Zsummary, add threshold lines
  if (p==2)
  {
    abline(h=0)
    abline(h=2, col = "blue", lty = 2)
    abline(h=10, col = "darkgreen", lty = 2)
  }
}


#
######INTERESTING MODULES
# Genenames
geneNames <- readRDS("../Data/EnsDb.Hsapiens.v75.Rds")

#select modules
#In Control
intModules_control = c("darkorange", "darkgrey", "orange", "pink", "black")
probes_control = colnames(counts_PW_control)
probes2annotation_control = match(probes_control, geneNames$gene_id)
probes_control = geneNames$gene_name[probes2annotation_control]

for (module in intModules_control) {
  #Select module probes
  modGenes = (mergedCOlors_control == module)
  #Get their name
  modGeneNames = probes_control[modGenes]
  #Write to file
  # Write them into a file
  fileName = paste("GeneNames- NewControl -", module, ".txt", sep="");
  write.table(as.data.frame(modGeneNames), file = fileName,
              row.names = FALSE, col.names = FALSE)

}

#In case
intModules_case = c("pink","black")
probes_case = colnames(counts_PW_case)
probes2annotation_case = match(probes_case, geneNames$gene_id)
probes_case = geneNames$gene_name[probes2annotation_case]

for (module in intModules_case) {
  #Select module probes
  modGenes = (mergedColors_case == module)
  #Get their name
  modGeneNames = probes_case[modGenes]
  #Write to file
  # Write them into a file
  fileName = paste("GeneNames- Case -", module, ".txt", sep="");
  write.table(as.data.frame(modGeneNames), file = fileName,
              row.names = FALSE, col.names = FALSE)

}


```

```
#Reading tables from consensus path DB
CPDB_black_control <- read.table("CPDB_black_control_geneName.tab", sep = "\t", header = TRUE)
CPDB_darkgrey_control <- read.table("CPDB_darkgrey_control_geneName.tab", sep = "\t", header =
    TRUE)
CPDB_darkorange_control <- read.table("CPDB_darkorange_control_geneName.tab", sep = "\t", header
     = TRUE)
CPDB_orange_control <- read.table("CPDB_orange_control_geneName.tab", sep = "\t", header = TRUE)
CPDB_pink_control <- read.table("CPDB_pink_control_geneName.tab", sep = "\t", header = TRUE)

CPDB_pink_case <- read.table("CPDB_pink_case_geneName.tab", sep = "\t", header = TRUE)
CPDB_black_case <- read.table("CPDB_black_case_geneNames.tab", sep = "\t", header = TRUE)


#Intersect 2 moduler
modTOM_control_top <- colnames(modTOM_control[top, top])
modTOM_case_top <- colnames(modTOM_case[top_case, top_case])

intersect(modTOM_case_top, modTOM_control_top)
```

Rfiles/Analysis.R

```
#
#CYTOSCAPE
#
probes_control = colnames(counts_PW_control)
probes_case = colnames(counts_PW_case)

modules_cyto_control = c("black")
modules_cyto_case = c("black")

inModule_control = is.finite(match(mergedCOlors_control, modules_cyto_control))
inModule_case = is.finite(match(mergedColors_case, modules_cyto_case))

modProbes_control = probes_control[inModule_control]
probes2annotation_control = match(modProbes_control, geneNames$gene_id)
probes_control = geneNames$gene_name[probes2annotation_control]

modProbes_case = probes_case[inModule_case]
probes2annotation_case = match(modProbes_case, geneNames$gene_id)
probes_case = geneNames$gene_name[probes2annotation_case]


modTOM_control = TOM_control[inModule_control, inModule_control]
modTOM_case = TOM_case[inModule_case, inModule_case]

dimnames(modTOM_control) = list(probes_control, probes_control)
dimnames(modTOM_case) = list(probes_case, probes_case)

nTop = 1000;
IMConn = softConnectivity(counts_PW_control[, modProbes_control]);
top = (rank(-IMConn) <= nTop)


IMConn_case = softConnectivity(counts_PW_case[, modProbes_case]);
top_case = (rank(-IMConn_case) <= nTop)



cyt_control_top = exportNetworkToCytoscape(modTOM_control[top, top],
                                       edgeFile = paste("CytoscapeInput- Control_top- edges-",
                                           paste(modules_cyto_control, collapse = "-"), ".txt",
                                            sep = ""),
                                       nodeFile = paste("CytoscapeInput- Control_top- nodes-",
                                           paste(modules_cyto_control, collapse = "-"), ".txt",
                                            sep = ""),
                                       weighted = TRUE,
                                       threshold = 0.1,
                                       nodeNames = modProbes_control,
```

```
                                        altNodeNames = probes_control,
                                        nodeAttr = mergedCOlors_control[inModule_control])


cyt_case_top = exportNetworkToCytoscape(modTOM_case[top_case, top_case],
                                      edgeFile = paste("CytoscapeInput- CaseTOP- edges-", paste(
                                          modules_cyto_case, collapse = "-"), ".txt", sep = ""),
                                      nodeFile = paste("CytoscapeInput- CAseTOP- nodes-", paste(
                                          modules_cyto_case, collapse = "-"), ".txt", sep = ""),
                                      weighted = TRUE,
                                      threshold = 0.57,
                                      nodeNames = probes_case,
                                      altNodeNames = probes_case,
                                      nodeAttr = mergedColors_case[inModule_case])
```

Rfiles/Cytoscape.R