Marie Gulla

# An integrated systems biology approach to investigate transcriptomic data of thyroid carcinoma

May 2019

Master's thesis

Master's thesis 2019 Marie Gulla

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

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

# An integrated systems biology approach to investigate transcriptomic data of thyroid carcinoma

Marie Gulla

Biotechnology MBIOT5
Submission date: May 2019
Supervisor: Eivind Almaas
Co-supervisor: André Voigt

Norwegian University of Science and Technology
Department of Biotechnology and Food Science

# An integrated systems biology approach to investigate transcriptomics data of thyroid carcinoma

Marie Gulla

May 15, 2019

*“ I can appreciate the beauty of a flower. At the same time, I see much more about the flower (..). I could imagine the cells in there, the complicated actions inside, which also have a beauty. I mean it’s not just beauty at this dimension, at one centimeter; there’s also beauty at smaller dimensions, the inner structure, also the processes. (...) The science knowledge only adds to the excitement, the mystery and the awe of a flower. It only adds. ”*

Richard Feynman

# Abstract

Driven by the development of innovative approaches to quantify gene expression levels across large numbers of samples, differential transcriptome analysis is emerging as a powerful strategy to interrogate the complex interplay of genes accountable for malignancies. The CSD method is a correlation-based method to systematically classify differential genetic associations, facilitating identification of dissimilar interactions driving pathogenesis. In this work, we have used the CSD framework for analyzing gene correlation for thyroid carcinoma (THCA) patients. THCA is the most common endocrine cancer type. These tumours frequently resist standard treatments and are thus associated with poor clinical outcome. By using publicly available samples from The Cancer Genome Atlas, the transcriptomic landscape was investigated by contrasting these to normal thyroid expression profiles. The CSD method successfully pinpointed several interesting gene pairs in networks enriched for processes linked to carcinogenic behaviour. Examination of gene interactions revealed relevant gene groups driving aberrant signaling and regulatory cascades. Looking into well connected network regions identified hubs coordinating destructive information processing, likely responsible for deteriorated mechanisms needed to combat tumor progression. Probing gene associations characterized by transition into abnormal character resulted in potential novel prognostic markers of thyroid carcinoma.

In the second part, robustness and potential method improvements to the CSD framework were assessed. Quality control investigation demonstrated that obtaining consistent analysis results required proper data pre-processing, including batch effect correction. A fundamental step in correlation-based methods for differential studies, is quantifying gene-pair relationships from gene expression data. Here, we explored three alternatives to the conventional inference algorithm. First, weighted topological overlap (wTO) with soft thresholding was applied. This provided a robust computation, also giving meaningful results in the case of low sample sizes and appeared to produce biologically meaningful modular structures. The second method was based on computing the mutual information (MI) as a more far-reaching similarity measurement. Although it was more dependent on larger sample sizes, it elucidated numerous novel relevant gene pairs not captured by Spearman or wTO. Motivated by achieving a computational reduced footprint allowing applicability to larger data sets, the last alternative involved a simplified version of CSD omitting variance estimation. While maybe offering some false positives, the relaxed condition will produce useful result sets even for very large transcriptomic data. For quality assessment, gene interactions identified by any of the similarity measures, were analyzed with regard to biological function and significance. Alternative similarity measures augment the outcomes of the original CSD method, and yield candidate genes which may contribute to deciphering the pathogenesis of THCA.

# Sammendrag

Som følge av store framskritt i teknologi og innovativ tenkning har det blitt mulig å måle uttrykk av gener og hvordan dette uttrykket varierer over store grupper mennesker. Sammenligning av genuttrykk gjør det mulig å forske på komplekse sammenhenger mellom genuttrykk og sykdom. CSD-metoden er et rammeverk for å undersøke systematiske forskjeller i genetiske interaksjoner som er nyttig for å identifisere forskjellene som forårsaker sykdomsutvikling. I første del av denne oppgåven har vi sett på genuttrykk målt i skjoldbruskjertelen ved kreft. Skjortbrukskjertelkreft er den mest forekommende krefttypen blant de endokrine kjertlene. Det er en vanskelig kreft å bekjempe med nåværende behandlinger, og fører derfor ofte med seg dårlige sykdomsprognoser. På The Cancer Genome Atlas be genuttrykksprøver lasted ned og benyttet for å studere genuttrykksprofilen for denne krefttypen ved å sammenligne disse med prøver fra sunt vev. Med CSD-metoden greide vi å finne mange interessante genetiske korrelasjoner som framstilte et system med overpresentasjon for kreftrelaterte prosesser. Ved å se nærmere på disse, identifiserte vi potensielle kilder til feil i reguleringen av genuttrykk, for bland annet gener som er viktige i å koordinere ulike cellulære funksjoner som syntes å oppføre seg unormalt. Komparativ analyse av genuttrykk resulterte i nye kandidatgener som markører for skjoldbruskkjertelkreft.

I den andre delen av oppgaven forsøkte vi å undersøke robustheten til CSD-metoden og finne mulige forbedringspotensialer. Kvalitetskontroll, inkludert korreksjon for kilder til forstyrrelser bland målinger i datasettene, viste seg å være viktig for å forsikre pålitelighet blant resultatene. En grunnleggende del av analysemetoder basert på samenfallende mønster i genuttrykk er å beregne korrelasjoner mellom uttykksmålingene. Her ble tre nye alternativer for å gjøre dette utforsket. Først ble vektet topologisk overlapp benyttet, og resulterte i et robust likhetsmål også for mindre datasett. Dette var spesielt nyttig for å finne interessante grupperinger av korrelerte gener. Den andre metoden var å beregne gjensidig informasjon, og besto i å implementere estimering av entropi. Denne viste seg å være mer avhengig av tilstrekkelig antall målepunkter men bidro til å belyse nye viktige genetiske interaksjoner som ikke de tidligere fremgangsmåtene fanget opp. En siste alternativ metode var en forenklet versjon av CSD-metoden, motivert av redusert beregningstid for anvendelse på store mengder genuttrykksmålinger. Dog denne lettere framgangsmetoden kan være preget av noen false positiver, vil den være nyttig i anvendelse på store sett med genuttrykksmålinger. For å evaluere kvaliteten til de utviklede metodene så vi på deres evne til å fremheve biologisk relevante genetiske interaksjoner. Alternative likhetsmål som utvidelser til CSD-metoden økte kunnskapen om skjoldbruskkjertelkreft og dro fram nye geninteraksjoner som kan bidra til å forstå sykdommen bedre samt å utvikle nye behandlinger i fremtiden.

# Preface

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

I would first like to express my gratitude to my supervisor, professor Eivind Almaas, for his encouragement and valuable help throughout this work, in the form of shared insights, challenges, and feedback. Professor Eivind Almaas introduced me into this area and provided excellent supervision throughout this project has been indispensable. To my co-advisor André Voigt I will express my sincere gratitude for taking great interest in my project, teaching me linux commands, debugging, and for all fruitful comments on my work. He offered invaluable guidance, explanations of the CSD framework, and explaining whatever questions I might have. Thanks to both for constantly being available and responding to questions even in weekends and late evenings and giving me comment on this thesis. I would also like to acknowledge the support and encouragement from the fellow students of the Network Systems biology group, especially Martina, Emil, and Snorre. It has been a great team to be a part of.

I would also like to thank some fellow students for setting aside time. To Håkon Hukkelås for answering questions related to Python programming, to Jakob Pettersen for all help in working out programming challenges, and in particular employing the magic of parallel R. Madelene, I look up to you and thank you deeply for being my best friend and for the encouraging post-it notes you leave for me in our apartment. Thank you to my best friends Magnhild, Åsa, Victoria, Ragnhild, and Marianne. You have made these past five years my best. To all my friends and classmates who have provided a fun and inspiring work environment. Thanks also to the bad spring weather in Trondheim this year, making it possible to complete this work without being too annoyed about sacrificed outdoor activities.

Lastly I want to thank my parents and family for all support and encouragement, especially by brother Jan for teaching me Latex. To my dog Simba for believing in me. And last but not least, a warm thanks to my boyfriend Matias, for his love, patience and for lightening up my everyday.

Marie Gulla
May 2019

# Contents

# List of figures

# List of tables

# Nomenclature and notation

| **Abbreviations** | |
|---|---|
| $DEG$ | Differentially expressed gene |
| $DCG$ | Differentially co-expressed gene |
| $DGC$ | Differential gene co-expression network |
| $DGCN$ | Differential gene co-expression network |
| $GO$ | Gene ontology |
| $THCA$ | Thyroid carcinoma |
| $C, S, D$ | Conserved, Specific, Differentiated |
| $H$ | Node homogeneity in link type distribution |
| $k$ | Node degree, equal to number of links attached to the node |
| $p$ | Importance level |
| $k_p^{C,S,D}$ | Threshold value for C-, S-, and D-score in network inference |
| $\rho$ | Spearman's ranked correlation coefficient |

| **Comparison of similarity measures** | |
|---|---|
| $CSD$ | Conventional CSD-framework with basis in $\rho$ |
| $CSD\text{-}VAR$ | CSD with $\rho$ without correction for internal variance in $\rho$ |
| $wTO$ | CSD based on weighted topological overlap transformation of $\rho$ |
| $MI$ | CSD based on the mutual information |

# Chapter 1

# Introduction

## 1.1 Complexity of Biological Systems

Cells can be best understood as complex systems. Each of the thirty-seven trillion cells that make up the human body are complex machines utilizing elaborate schemes of orchestrated genetic transcription to create a self-maintainable structure that responds to its surroundings and contributes to the performance of a multi-cellular living organism. On the single cell level - as well as the multi-cellular level - integration of information ensures correct cellular conduct in different situations. Internal information is stored in the DNA, which is the blueprint for making proteins and other functional molecules necessary for the survival of the cell. Thousands of nucleotides constituting the genetic code are copied and translated into proteins each minute, and even at these astounding velocities it happens virtually without errors. External information is received through its membrane and the proper signalling responses are mediated through extremely speedy transduction pathways transmitting information through the cell to the fitting target. Before a second has passed, thousands of molecules have been synthesized and complexes have been assembled by elaborate molecular factories. Matter flows in and out of the cell as it fuels its exquisite show of complex life in nature's most minuscule entity.

But the cell is also matter that dances. Perhaps the ultimate characteristic of a living substance, is that it does not obey precise laws but its activities are associated with noise. Stochastic fluctuations within a cell makes it phenotypically different from its surrounding cells, even though they may share identical DNA. Cells are under strong thermal noise, and in the dense soup of molecules that fills the cell macro-molecular complexes spontaneously assemble to perform a task and dissolve and vanish without effort when the work is done. The existence of randomness reflects the intrinsic complexity of the cell. Complexity in biology does not have any operational definition and cannot be captured by any common measure, because it refers to all structural, functional and hierarchical complexity [6].

In spite of this, scientists have attempted to develop generalizable principles on the conduct

of living systems. Living systems were first studied by breaking them down to manageable pieces that could be understood, but this changed during the 20th century with scientific emphasis on whole indivisible behaviour. It became evident that the structure of the entire system as a whole orchestrated the behaviour of its components. This lead to the development of high-throughput experimental technology providing the necessary level of comprehensive and detailed information [7]. Because biology has evolved to perform specific tasks, it became apparent that the style of biological models was not haphazard but indeed describing a function. Resulting from the last 30 years of growth in processing power, storage capacity and interdisciplinary collaboration; systems biology emerged as a quantitative integration of interacting molecular components in a dynamic system made possible by increasing computational power. System biology models the converging patterns of cellular circuitry that obey some framework enough for it to be expressed mathematically and the recognition of networks in biology [8].

Integrative thinking is the foundation for the field of network biology. Biology has slowly been unraveled by combining mathematical tools with high-quality experimental data to construct and simulate data-driven networks. Networks are employed to model complex biological systems because it captures the complexity of the system and facilitates detailed analysis of the its structure and the characteristics of its components in a stable yet dynamic manner.

Most cells must constantly monitor and fine-tune the genes it expresses and the levels of genetic transcript it produces. The genetic products - the RNA molecules and proteins - perform tasks that are both necessary for the ordinary cellular life and to respond to unforeseeable obstacles and challenges. During its entire life cycle, the cell needs to synthesize and do maintenance of its building blocks, manage the quality of its machinery, and make sure there is constant flow through its energy-producing metabolism by both taking up nutrients and flushing out waste products. The model properties representing different cell states are determined by the gene-expression profiles, indicating which genes are actively being transcribed and to which degree. The large number of cell states and their reproducibility by mathematical models attest to the existence of molecular programs ensuring reliable execution [9]. It is also a description of the cells in their true functional states, not their functional capacity, given merely by their genetic make-up.

In contrast, gene expression analysis illuminates the cells accurate phenotypic behaviour. Cells of the human body contain around 20 thousand genes. The 1000 Genomes Project Consortium found that 99.4% of the genome in humans is the same [10]. Yet there are millions of differences between any two people. Phenotypes are variations in the outside appearance resulting from many factors within the genetic makeup; like polymorphisms and variations in gene expression levels. Investigating variances between people has lead to major discoveries in medicine and continues to help us learn more about human health and disease. From a gene in the DNA is expressed in the nucleus of a human cell, transcribed into RNA, processed into messenger RNA (mRNA), transported and translated (if protein-coding) into amino acids and folded into functional proteins, there are several potential things that may

go wrong and several sites of regulatory interacting ensuring genetic expression of optimal levels. Through evolution, robust machinery have evolved that control the quantity and quality of information flowing from the genetic material into expressed genes and hinder any erroneous events within the process of genetic expression from permeating into the cellular system. Small fluctuations within this cellular quality-control machinery give rise to phenotype variation. However, some errors might occur and slip through the machinery as well. Incorrect level of genetic expression from a normal gene, or expression of a gene with mutations, could ultimately lead to malignancies in the cells [11].

Gene expression analysis have been extremely useful for investigation of human phenotypes and diseases. Analysis of variation in this fine-tuned flow of genetic information aims to deduce how aberrant gene expression and disruption in the regulatory machinery may lead to malignancies. Many human diseases are manifestations of disordered genetic interplay in specific tissues. Genes that may cause diseases would be expressed at abnormal levels in the tissues where these defects cause pathogenesis. The underlying complexity of genetic regulation is highly embedded in natural regulation of the entire organism [12]. Gene expression analysis investigates correlations among expression levels between genes and is thus a great tool to examined genetic interplay in detail. As quantitative measures, gene expression levels are well suited for co-expression analysis, where their correlated expression levels characterize the underlying patterns of transcriptional regulation. Juxtaposition of co-expression analyses from gene expression measurements originating from cells of different nature, enables comparison of the regulatory schemes and expression patterns descriptive of each condition and facilitates contrasting properties from each condition. Comparing co-expression profiles based on gene expression measurements from healthy and sick persons thus allows for the investigation of abnormal events related to transcription, and which differentiated behaviour is likely to drive damaging processes to the system. A systems' biology approach thus provides insight into disease-related processes and their characteristics. This knowledge may guide better therapeutic approaches and the development of predictive and preventative medicine [13].

## 1.2 Aims and objectives of this thesis

This thesis has two underlying goals that are complementary to each other. The first is to get acquainted with and employ the established CSD-framework to perform a detailed differential co-expression analysis. This part focused on using gene expression measurements obtained from The Cancer Genome Atlas of patients diagnosed with thyroid carcinoma to contrast underlying cellular processes by comparing the co-expression correlations with those of normal persons. The objective was to perform an in depth investigation of central players and the underlying mechanisms these were involved in characteristic of thyroid cancer. In addition, network analysis tools were employed to identify network neighborhoods that could represent disease modules which could reveal novel patterns driving the pathogenicity within the transcriptomic system in thyroid cancer.

The second part aimed to contribute to the the CSD-framework with method development. This section aimed to integrate several tools of bioinformatics and systems biology in a congruent pipeline to produce reliable information from experiments across different data bases. Expanding the range of alternative similarity measures forming the basis of differential gene co-expression analysis was an intriguing quest which motivated development of efficient ways to incorporate new similarity measures into the CSD-framework. This pipeline needs to manage data-sets across cohorts, formats and programming languages. The influence these had on the results needed to be examined as well, in order to evaluate which potential benefits to the original CSD-framework these provided.

This master thesis has two major research goals that will be reflected in the presented work and the organization of the thesis paper. These are as follows:

1. Perform a comprehensive study of differentially co-expressed genes between healthy individuals and those that suffer from thyroid carcinoma, discover relevant gene modules, and explore interesting co-expressed genes that represent novel candidates for prognostic genes of thyroid cancer.

2. Integrate several tools to improve the current CSD-framework and inspect the quality of those. I intent to develop work-flows for application of alternative similarity measures and to evaluate their limitations and potentially important benefits.

# Chapter 2

# Theoretical background

This chapter will introduce the topics investigated and the theoretical foundations of methods utilized for data analysis in this thesis. Most of the topics explained in this section will provide the basis for understanding the strategies employed - especially for performing differential gene co-expression analysis, which is the main agenda of this master thesis.

Concepts from systems and network biology are mostly obtained from two books, *Network science* by Albert Lázló-Barabási [1] and *A first course in systems biology* by Eberhard o. Voit [2]. The reader is encouraged to look into these works for more detailed information. The CSD method for inferring differential gene co-expression networks developed by Voigt et. al. [3] is the first description of this method and the source of the information provided about it here.

## 2.1 Thyroid cancer

Thyroid cancer is a common endocrine malignancy, which has one of the highest increases in incidence globally among all cancer types [14]. Thyroid cancer incidence has persistently increased on a global scale. Papillary thyroid carcinoma is reported to be the most common type of cancer to have a documented increased incidence rate [15]. The most common histologically different thyroid cancers originate in either follicular or parafollicular cells. Papillary thyroid carcinoma (PTC) and follicular thyroid cancer (FTC) are the two types of well-differentiated thyroid cancers. PTC and FTC has the highest prevalence among the thyroid cancers [16]. PTC lesions are quite varied and can occur anywhere in the thyroid gland. Typical lesions are 2-3 cm in average size, are firm and have an invasive appearance. The nuclei of cells with PTC are typically clear and oval, are larger and contain more hypodense chromatin than normal and overlap with another [17]. The cancer tumours may invade the lymphatic system and lead to multifocal lesions and regional lymph node metastases [18]. In the cases where the thyroid cancer invades the vascular system, it is associated with a more aggressive disease and higher incidence of recurrence [19]. FTC

differs from PTC by not sharing these nuclear features, and is typically characterized by presence of capsular or vascular invasions. FTC is more likely to spread to other tissues because it more readily invades blood vessels [20].

Among the differentiated thyroid cancer types, papillary thyroid carcinoma has increased the most, approximately 3-fold between 1988 and 2002. Almost all of this rising incidence consisted of cancers measuring 2 cm or smaller [21]. The treatment of differentiated thyroid cancer normally includes risk assessment by neck imaging followed by surgery to remove tumor tissue. For non-metastatic tumor nodes, a lobectomy has been shown to be associated with long-term survival [14].

Genetic studies of thyroid cancer have shown that many types harbour genetic changes affecting signaling pathways, which shift the correct regulation of growth and cell proliferation [22]. Several factors affect the risk of developing thyroid cancer, such as dietary iodine and exposure to radiation. Radiation has been shown to induce apoptosis, cell cycle arrest and cancer in the thyroid gland [23]. Thyroid cancer, as well as other thyroid endocrine diseases, may arise from radiation-induced damage to thyroid follicles. Nuclear bombs have caused thyroid cancer in survivors of the bombing in Japan during World War II. After the nuclear plant explosion in the former USSR nuclear radiation lead to thyroid cancer *in utero* and in small children.

The tumor suppressor genes p53, encodes a family of proteins involved in thyroid cancer development, and are targets for novel therapeutic strategies [24, 25] Recent research has also shown that some specific genetic variations are related to an early onset of autoimmune thyroid diseases, indicating that the immune systems plays an important role in thyroid disease pathology[26].

Tumor gene expression profiling brings new insight into cancer pathophysiology and contributes to better understanding the molecular basis of the malignancy. Often, many functions in the cancer cells are similar, if not equal, to normal cells of the same tissue. When most cellular properties are retained, discriminating among cancerous and normal tissue becomes a difficult task. This makes both discovery and proper diagnosis of the cancer challenging. Thyroid cancer, especially, is expected to be less effectively distinguished from non-transformed tissue compared to other malignant tumours because its gene expression profile lacks obvious carcinogetic characteristics [27]. An additional complicating aspect of thyroid carcinoma is that carcinogetic cells are usually intermingled with normal cells, so that quantitative associations may vary even inside the same tumour tissue. Although thyroid carcinoma is clinically heterogeneous, some studies of micro-array expression profiles have found distinct clustered profiles [28], motivating further studies of the characteristic thyroid cancer expression profile.

## 2.2 Microarray technology

Cells constantly need to regulate their gene expression levels in order to meet the demands of an ever-changing environment. Different challenges are met by transcription of a gene who's gene product has the desired function. The cell responds to external signals by letting information flow from its peripheral signal receptors into its nucleus, where a transcriptional regulator mediates the appropriate is translated into a cellular response. This genetic response is in the form of messenger RNA (mRNA). The genetic transcripts leave the nucleus so that they may perform various tasks in the cell, either as they are or after being translated into protein [29]. To ensure optimal quality and functionality of the cell's machinery, specific genes are transcribed so that new molecules can replace old ones and the correct amount of each gene product is available. The transcripts thus regulate which processes are taking place and tells us a great deal about what the cell is responding to [30].

The motive for developing the microarray was to create a technology that could take a snapshot of the transcriptional status in a cell at a given time to frame which mRNAs have been made and at which levels. The aim of the assay is to quantitatively evaluate RNA abundances. Because they are key intermediates between genes and gene products, they are very informative. This is referred to as the cell's *transcriptome* and is measured by reverse transcriptions of cell sample mRNAs into cDNA molecules with the simultaneous incorporation of marker molecules - either with radioactive or fluorescent labels. These are applied to a microarray, a membrane on which probes for the DNA-sequences of the organism are attached. These probes are immobilized at fixed positions and allow multiple sample cDNA molecules from the cell to hybridize to each probe, for quantitative measures for each gene of interest. Sample cDNA only hybridizes to the array fragment if the sequences are complementary, and detection of label intensity for each probe spot thus makes this technology an accurate assay of gene transcript quantification for living cells. These arrays are commonly used to measure the gene-expression patterns of human cells from various sick tissues. The information can thus be used to study the transcriptomic manifestations of diseases [31, 32].

## 2.3 RNA sequencing

High-throughput screening of patients is a process generating large quantities of information of diverse types. It is employed in both basic and applied research to facilitate analysis of genes and their function, global genomic expression and regulation [33]. For an integrated data analysis, data of good quality is needed. Hybridization-based micro-arrays are effective in providing an expression snapshot of known genes and have been a robust and reliable method for expression analysis for many years. They are, however, insensitive to splice isoforms and are less suited to discover unknown genes, and may thus not capture genes that may in fact be significantly associated with a specific phenotype. The development of

next-generation sequencing provided a deeper and more quantitative view of gene expression, alternative splicing and allele-specific expression [34]. Gene co-expression networks can be reconstructed from gene expression data from micro-array or RNA-sequencing technology. RNA-sequencing enables more comprehensive analysis of the transcriptome and has thus become widely used for generation of gene expression data in recent times [35].

Next-generation sequencing of mRNA is termed RNA sequencing (RNA-seq) and is a technique for detailed transcript identification and quantification. RNA-seq data reflect a biological systems repertoire of RNA molecules at a given time. For the well-annotated genome of *Homo Sapiens*, all transcripts may be mapped to a genetic region directly. The number of reads mapping to the target gene are measured with the *read count*, which is linearly related to the transcript abundance [36]. The method typically consists of isolation of RNA extract from a cell or tissue population, converting the mRNA to cDNA with reverse transcriptases, preparing and sequencing this cDNA-library on a sequencing platform. RNA-seq relies on short sequenced nucleotide reads mapping to a genome assembly. Illumina has dominated the sequencing industry with its sequencing by synthesis approach, in which clonal amplification of molecules enable quantitative measures of expression with very low sequencing error rates. Illumina sequencing has advanced greatly and is now able to sequence over 500 base pairs from the same mRNA-fragment [35]. Compared to micro-arrays, it has been demonstrated to identify multiple more true positive differentially expressed genes between different tissues[33].

RNA-seq offer many beneficial aspects to the expression technology. One of the biggest advantages by using RNA-seq it that the expression of a huge amount of genes can be quantified, even some 70 thousand non-coding RNAs, recently annotated genes encoding long interfering non-coding RNA-molecules (lincRNAs). Because lincRNAs may have a role in human diseases [37] and cancer [38], using RNA-seq data offer a possibility to include these in the analysis. Moreover, RNA sequencing provide major benefits for genetic accuracy. This technology has very high resolution and higher accuracy for transcripts of low-abundancy [39]. It can distinguish genetic paralogues [40] and detect detect structural variations in the genome that was not previously possible, like genetic fusions and alternative mRNA splicing events [41]. However, determination of expression levels for different splice variants of the same mapped exon may remain challenging, limiting the co-expression analysis on splice-variant level. To perform co-expression analysis with splice variants, one needs to distinguish all possible splice variants for one gene instead of mapping these together. While this method preserves information about the co-expression of different transcripts encoded by the same gene, the size of the square co-expression data matrix and subsequently the needed computational capacity drastically increases. The most common practise for RNA-seq co-expression networks is therefore to merge the expression data for overlapping gene isoforms and construct a co-expression network at genetic resolution [11]. Co-expression network reconstruction from RNA-seq data have higher functional connectivity if there are enough samples, and attain higher quality compared to micro-arrays with same number of samples mainly when the sequencing depth cut-off is 10 million reads per sample [42].

## 2.4 RNA-seq data processing

Gene expression quantification data obtained with RNA-seq needs further bio-informatic processing. Data needs to be both aligned and assembled. Firstly, the mRNA expression data needs to be aligned to the genome of the organism of which the gene expression is measured. All nucleotide sequences of the mRNA raw reads generated with RNA-seq are compared to the sequence of the annotated genome and mapped to a region if the sequences are identical. This process converts unaligned reads to all aligned reads [43].

Quality control is performed to assess potential issues in the expression data. To mention a few, FastQC, RSeQC, and QoRTs are all excellent tools for QC that detect possible defects or low quality in the read data. FastQC has a extensive range of control measures, and may also be used to obtain data on the prevalence of splicing effects [44]. Data is aligned to a reference genome for a given organism in the form of a GTF-file. These files detailed information on the genetic material the organism expresses, like its position in the genome and which strand of the DNA-molecule it resides on. One of the most widely used tools for RNA-seq read mapping is STAR [45]. This tool is compiled with a gcc c++ compiler. It uses a large amount of memory but can map reads around 50 times faster than other equivalent programs. Also, it is readily used to detect novel splice variants [11].

Second, after raw read alignment the mapped reads need to be quantified at each genetic region. Aligned genomic data files, typically .bam- of .sam-files, are further processed with read-counting tools to produce a precise quantification of expression levels at each genetic loci. Gene expression quantification data requires a suitable set of data processing tools [43]. For quantification HTseq [46] is a very common tool. It is a Python framework for high-throughput transcriptome profiling data. This tool assigns processed read counts based on aligned reads from STAR, and produces RNA-seq expression level quantification data. A great benefit of using HTseq is the application of its pre-processing tool which is especially suited for differential expression calling. In this case, the reads are not normalized, because normalized transcript values should be used only within the context of the entire genetic set and not across experiments or cohorts [47]. Instead, expression levels are in counts-format, which is the raw number of reads aligned to each gene. This makes expression data in counts-format well suited for subsequent differential analysis [47].

### 2.4.1 Normalization methods and expression level modelling

RNA-seq data is made up up gene counts aligned to a genome which needs to be corrected for possible sources of measurement biases. These can be positive correlations between gene size and read count and variable sequencing depths [48]. Read counts sampled from a fixed set of genes would follow a multinomial distribution, originally approximated by the Poisson distribution [49] but has recently been replaced by the negative binomial distribution to address the underestimation of variation with Poisson [50].

Count-based methods for modeling expression data with the purpose of identifying differentially expressed genes is challenged by potential dependence between samples and over-dispersion. Expression data generated with RNA-seq have been modeled with Poisson models, which is appropriate for technical replicates. In many cases there is over-dispersion present in the data, which has lead to the frequent use of the negative binomial model for gene expression patterns [51].

## 2.5 Network theory

Network science is the most important theoretic foundation for understanding models of complex biological systems. Modelling biology with networks is a structural and quantitative representation of the system. These computational models may describe processes at all cellular levels; gene regulation, signal transduction, entire energetic metabolism and even complete tissues or whole organisms [52]. Perturbing the network to simulate various challenges the system may encounter shows a great deal about its properties that may only be elucidated when the system is described as a whole. Visualizing the complex system as a network graph is usually done by giving each component connections to each other based on their level of interaction, each component is represented with a node and their connections are edges between them [1].

Yet, most important for the success of employing networks in biology, is that it enables mathematical modeling and computations on a system too large to draw and inspect with one's eyes. In the field of graph theory, many tools have been developed to extract important properties of networks, so that even those containing thousands and billions of connections can be investigated. Characterization of network properties and the effect these have on the biological system may elucidate aspects of it not available through other forms of analysis. Network science has thus become invaluable in the desire of identifying the true governing principles of complex biological systems in nature [2].

The principal representation of a network is with a *graph*, in which nodes may be connected by *edges* and edges may exist only between a set of nodes. A network typically consists of a collection of interacting elements tied together by the elements' pair-wise interconnectivity. The set of items in the networked system represent the functional components and edges between them represent an existing connection between these components. When representing a dynamic system as a network its structures are defined by the components, the *nodes*, independent of their interpretation. The number of nodes $N$ in the network is referred to as its size. The number of edges in the network, $L$, is the total number of interconnections between the nodes [1]. Fig. 2.1 shows and example of a simple network with four nodes connected by four links.

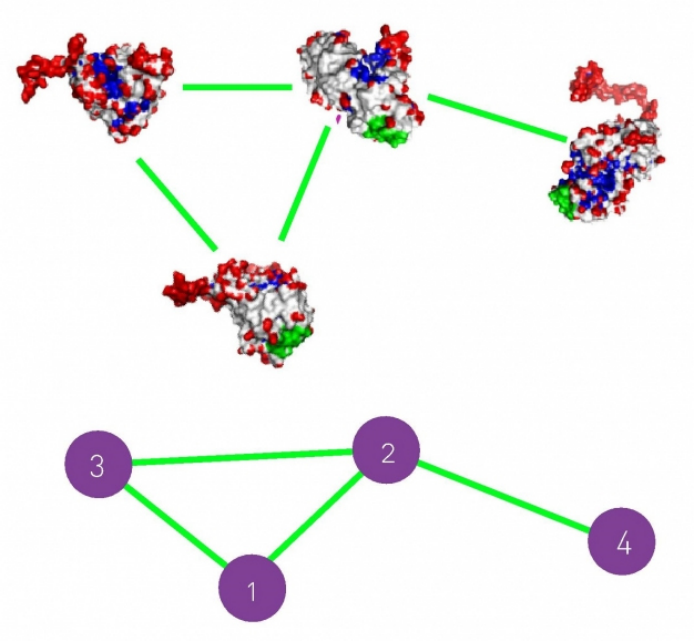


**Figure 2.1:** The figure shows a small network of four nodes and four links, in which the the top one shows four proteins and their connections and the bottom is a simple graph representation of the network. Taken from [1].

Random graphs are graphs artificially created by making a set of nodes and randomly assigning edges between them based on a probability of connection $p$. Such random network models were first studied by mathematicians Pál Erdős and Alfréd Rényi and are called *Erdős-Rényi networks*. Many forms of network characterization are performed by comparing a real network to a artificial network created with a random network model of a similar size[1].

### 2.5.1 Adjacency matrix

In its simplest form, a network can be summarized in an *adjacency matrix* if it is fully specified. The adjacency matrix is a two-dimensional matrix given by all its components and contains the attribute of all pairwise combinations. In the adjacency matrix $A_{ij}$, each element $a_{ij}$ quantifies a connectivity between the components $i$ and $j$. The elements in the matrix hold information about whether or not there is a connection between the elements, thus represented as an edge in the network.

For an *undirected network*, $A_{ij}$ is symmetric, e.g. $a_{ij} = a_{ji}$. The opposite is true for *directed networks*, in which the edges between nodes are directional and $A_{ij}$ may be unsymmetrical. Networks can differ in an additional way, they can take on either binary values or weighted values. A *weighted network* have weights assigned on it's edges, in this case the elements of the adjacency matrix may take any value. A common procedure when constructing networks from a weighted adjacency matrix is to use a hard threshold, as in Eqn. 2.23. This transforms a weighted $A_{ij}$ into a matrix with binary values describing whether the element $a_{ij}$ is over (1) or under (0) the required threshold. Conversely, the matrix elements

of *unweighted networks* have either 0 or 1 as values. The concentrated way of representing a system permits many different methods of linear algebra analysis. In this thesis, networks will be generated as results from the method development and application of software to the data under study. Unless otherwise stated, networks will be unweighted and undirected.

In biological networks, *nodes* in the network represent molecular or biological entities and are connected by *edges* if they have a common property or a shared connection. These connections link together nodes if they interact in some way, like molecular interactions affecting each other's properties and activity, enzymes responding to the same signal or chemical bonds tying together molecules into large complexes. The network is a useful and versatile tool that can describe a wide variety of structures, and becomes a map describing behaviour in different situations and the interplay of its components [1].

Network representation of complex systems have been applied to a wide variety of disciplines. Everything from gene regulation, protein-protein interactions and kinetics, the complete metabolism of various organisms, cell-to-cell signalling in all ranges of single-celled networks to multi-cellular tissue networks. By reconstructing the biological system in a network, it's connectivity and structure can be used to answer questions about ability, robustness and the course of information flow between its components. The representation of a dynamic biological system in a static graph is advantageous because of its degree of simplicity compared to a temporally changing network of fully regulated dynamic systems. Analysis of the biological network can then be performed by applying graph theoretic mathematics to analyse the system's structural features [53].

The elements of the adjacency matrix can represent many different entities in different settings, be it genes, micro-RNAs, proteins or neurons in the brain. A common assumption is that the collection of elements within any network representation are of identical type in respect to the system. Homogeneous elements represents a common format of information in the system, encoding various types of interactions. These can be co-expressed genes, transcriptional regulators, enzyme complexes or synaptic links. In many cases, it is therefore practical to study these independently on a detailed level. But networks may also consist of multiple such layers, in order to create a fuller description of information flow in a biological system.

### 2.5.2 Node degree

An important characteristic of a node is its degree, denoted $k_i$. The degree represents the number of associated edges connected to the given node in undirected networks and oppositely, in directed networks, the degree of a node will be defined as either in-degree or out-degree [2].

The average degree in a network is defined as

$$\langle k \rangle = \sum_{k=0}^{\inf} kP(k) \tag{2.1}$$

where $p_k$ is the proportion of nodes with degree $k$, given by the degree distribution, see equation 2.2 below.

### 2.5.3 Degree distributions

The distribution of the node degrees in a network is an important characteristic [2]. The degree distribution *P(k)* is a measure of node degree value frequencies in the network as a whole. *P(k)* is given by

$$P(k) = m_k/m \tag{2.2}$$

where $m$ is the total number of nodes and $m_k$ is the number of nodes with degree $k$. It is a function that shows the proportion of nodes having degrees on a scale of degrees $k$, from zero to the highest possible degree of that network. The sum of all values of *P(k)* equals one. This distribution thus describes the probability that any random node will have degree equal to $k$ [1]. For random graphs the degree distributions usually has a narrow bell shape, given by its binomial degree distribution with low degree of variance. In this case, the majority of nodes in the graph have a degree that is close to the average node degree [2].

The exact degree distribution of Erdős-Rényi (random) networks,

$$p(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k} \tag{2.3}$$

depends on the probability that $k$ links are present, $p^k$, the probability that the remaining $(N-1-k)$ links are not present, $(1-p)^{N-1-k}$, and the total number of combinations a number of $k$ links may be selected from the total $N-1$ links a node can have, $\binom{N-1}{k}$. Because real networks are sparse, the average degree is much smaller than number of nodes, so random network models approximate this trait by the Poisson distribution for when $\langle k \rangle << N$. This degree distribution is given by $p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$, and only depends on the parameter $\langle k \rangle$, which makes it a preferred formula for many calculations [1].

In networks describing real biological systems, it has been found that the degree distribution vary widely. It may follow a straight line when plotted on a logarithmic scale. This describes a system in which a very limited number of nodes have a large degree, few have a degree close to the average, and most will have very low degrees. Here the *P(k)* follows a *power-law distribution*,

$$P(k) = k^{-\gamma} \tag{2.4}$$

in which the slope $-\gamma$ is negative. A network whose degree distribution show these charac-

teristics is referred to as a *scale-free* network, because there exist no "typical" node degree and a common "scale" of degree among genes is lacking.

One interesting aspect of this general scale-free property in most kinds of real networks is that it demonstrates that biological networks are not organized randomly [3]. The scale-free property in real networks has been found to emerge as a result of a process called *preferential attachment*, which entails that the number of nodes is not fixed, and any new node is added will have links with preference to nodes that already have many neighbors [1]. This growth mechanism, which is well applicable to a genetic network are added by genetic duplication and functional divergence, will typically lead to a $\gamma$ that may range from small values towards infinity [3].

In many networks, both those representing biology and random graphs, there are numerous nodes with relatively low number of neighbors while some of them have many. Nodes with a disproportional huge number of links are termed *hubs*. The presence of hubs is one of the major traits that distinguish real scale-free networks from random networks. In fact, in a scale-free network the probability of a hub being present is several orders of magnitude higher than in a comparable random network [1].

### 2.5.4 Paths and betweenness centrality

Distances in a network are measures of how far away its components are from each other. The distance between two nodes, called *path length*, is a relative measure of how many links a path between the pair contains. There are often many alternative paths between two nodes in a network. The *shortest path* $d$ is defined as the path containing the fewest links that connect the nodes in a pair. In undirected networks, $d_{ij} = d_{ji}$. Oppositely, the longest path between any two nodes of a network is termed the *network diameter* [1].

*Betweeness* quantifies the number of shortest paths between node pairs of a network that go through a link (i, j). A link with high betweenness is a link through which a lot of information flows between nodes located further away from each other [1]. When looking at a specific node, the fraction of shortest paths that pass through it defines the *betweenness centrality* of the given node. It is defined as

$$C_B(n) = \sum_{ij}^{n} \frac{d_{ij,n}}{d_{ij}} \tag{2.5}$$

for all possible pairs of nodes $ij$ from all the $N$ nodes in the network, where $d_{ij}$ is the number of shortest paths between nodes $i$ and $j$ and $d_{ij,n}$ is the number of shortest paths going through node $n$, assuming $i \neq j \neq n$ [54]. Node betweenness centrality thus represents how important the node is for the spread of information through the network and reflect which role the node plays in controlling associations between distant nodes [55].

### 2.5.5 Eigenvector centrality

Like the betweenness centrality measures how important a node is for the shortest paths through a network, the *eigenvector centrality* measures a node's importance by accentuating nodes with connections to other important nodes. A node's eigenvector centrality is thus a measure of how many neighbors it is linked to, where each link to a neighbor is weighted by the centrality of this neighbor node.

Using the adjacency matrix $A = (a_{i,j})$, the eigenvector centrality $x_n$ of node $n$ can be computed. It is defined as:

$$x_n = \frac{1}{\lambda} \sum_k a_{k,i} x_k \tag{2.6}$$

where $\lambda \neq 0$ is a constant. In matrix form this is: $\lambda x = xA$. The eigenvector centrality vector $x$ is identical to the left-hand eigenvector of $A$ associated with the eigenvalue $\lambda$ [56].

The eigenvector centrality score for node $n$ is computed by determining the principal eigenvector of each node using the $m \times m$ adjacency matrix $A_{ij}$, given by [56]:

$$C_{EV}(n) = a_{1i}x_1 + a_{2i}x_2 + ... + a_{mi}x_1 \tag{2.7}$$

As this value is computed, if a node has a connection to a node of high eigenvector centrality score, this high-scoring node will contribute more than the low-scoring nodes to the eigenvector centrality of the first node. The resulting centrality measure a node's "influence" in the network [57].

### 2.5.6 Clustering

In a network the edges may be distributed differently between local regions. Computing and comparing properties of local and global edge patterns can provide meaningful information. These structural properties of one network neighborhood may vary greatly from other neighborhoods in the same network. Some neighborhoods are so tightly interconnected with themselves that they forms a clique almost entirely separated from the rest of the network. Quantification of such local structural properties of the *cliquishness* of a neighborhood is measured by the *clustering coefficient* [58].

The local clustering coefficient was first defined by Watts and Strogatz, and represent the likeliness that two network nodes which both share a common neighbor are also connected to each other, forming a closed triangle [59]. In a network neighborhood where a vertex (node) $n$ has a total of $k_n$ neighbors, there may be found a maximum of $(k_n(k_n-1)/2$ edges between these vertices. For a node $n$ in an undirected network, the clustering coefficient

$$C_n = \frac{k_n}{(k_n(k_n-1))/2} = \frac{2k_n}{k_n(k_n-1)} \tag{2.8}$$

where $k$ is the degree of node $n$, defined in $C_n \in [0, 1]$.

As long as there are a significant number of nodes with shared neighbors that are also connected to each other, the clustering coefficient for a node in this cluster will be non-zero and increase with the number of closed triangles it takes part in [60]. $C_n$ denotes the fraction of all possible links that actually exist in a specific local region of the network of all of the theoretically possible ones. If this maximum number of edges between nodes in a cluster really are present, all $n$ neighbors are connected to each other. In this case the clustering coefficient becomes one, because all links allowed in the neighborhood do exist. The clustering thus is a measure of the extent to which nodes are connected to each other in a closed clique [59].

For a complete undirected network, the *average clustering coefficient* is the average of all clustering coefficient of the network nodes,

$$C = \frac{1}{N}\sum_{i=1}^{N} C_n \tag{2.9}$$

where $N$ is the number of nodes in the network the clustering is averaged over [58]. The clustering coefficient will depend on interconnectedness in various cliques, and thus some will contain interconnected nodes with high clustering coefficients and others lower.

### 2.5.7 Network assortativity

Assortativity in a network is defined as the correlation of degree between connected nodes. When the majority of nodes of high degree tend to be connected to other nodes of high degree, the network has assortative properties. In assortative networks, hubs associate with other hubs and avoid low-degree nodes. If there is no correlation between the degree of a node and the degree of its neighbors, the network does not show any assortative characteristics. Conversely, if there is a tendency of hubs to *avoid* connecting to other hubs, so that the network mainly consist of hubs connected to many small-degree nodes, the network has dissassortative properties. For any given node to randomly choose the nodes it links to, the probability of that nodes of degrees $k$ and $k'$ in a network with $L$ links is given by

$$p_{k,k'} = \frac{kk'}{2L}. \tag{2.10}$$

Assortativity is thus a measure of how much the number of links between node $i$ of degree $k$ and node $j$ of degree $k'$ deviate from Eq. (2.6). The trend of node assortativity in the network as a whole is found in $e_{ij}$, the *degree correlation matrix*. For a randomly chosen link in the network the degree correlation denotes the probability of then observing a link between a node of degree $i$ and a node of degree $j$ [1].

This gives a *degree correlation function* and its approximation:

$$k_{n,n}(k) = \frac{1}{k_i}\sum_{j=1}^{N} A_{ij}k_j \approx ak^{\mu} \tag{2.11}$$

where $\mu < 0$ is a dissassortative network, $\mu = 0$ is a neutral network, and $\mu > 0$ is an assortative network.

## 2.6 Similarity measures

The elucidation of functional relationships is the overall goal of differential co-expression analysis of transcriptional data between conditions. From matrices of expression data, co-expressed genes are identified and grouped together. This bundling is produced by assigning relationships between genes based on a measure of similarity in their expression levels across multiple samples. The chosen similarity measure can have a large impact on the results of the analysis. Often genetic relationships may be measured with the Pearson or Spearman rank correlation measures. As an extension to these methods, similarity measures that are non-linear have also been employed in co-expression recently. The information-theoretic measure mutual information has been used to detect non-linear relationships between data sets that are not manifested in correlations [61]. Since this thesis later on exemplifies concepts of data analysis using gene expression measurements, genes are thought of as objects and may later be referred to as random variables of a data set.

The ability to infer real and meaningful relationships between genes in a complex network depends on the correlation detection methods used to calculate the degrees of associations between variables. Inherent to all co-expression networks, connections between genes must be based on a appropriate similarity measure. Based on the same set of expression data, network inference on the basis of different similarity measures may lead to very different genetic relationships [62]. Different similarity measures may elucidate various regularities in the transcriptomic data: simple positive correlations may identify patterns of co-regulated genes, correlation also including negative values may deduce patterns of antagonistically tuned processes, and mutual information may detect non-linear more complex genetic relationships between genes. Choosing an appropriate similarity measure thus becomes important for the success of the network inference [63].

Given two real-valued sequences of measurements $x = \{x_i : i = 1, ..., n\}$ and $y = \{y_i : i = 1, ..., n\}$ the similarity is a measure of dependency between the sequences. It is defined as a metric of similarity producing a higher value as the dependency between the compared variables increases if it satisfies the following criteria:

1. Limited range: $f(x, y) \leq 1$ for all $x, y \neq 0$
2. Reflexitivity: $f(x, y) = 1$ if an only if $x = y$
3. Symmetry: $f(x, y) = f(y, x)$

A similarity measure of 1 is the largest possible value representing perfect similarity between the variables, and a similarity of zero is the lowest possible value and represents the case of no similarity. Hence, they are directly applicable as values of the adjacency matrix $A_{ij}$

of which each element $a_{ij}$ hold information of associations between variables $i$ and $j$. In networks it is interesting to look at all pairwise similarities between all combinations of components, denoted in the *similarity matrix* $S = [f_{ij}]$. For a gene-expression matrix

$$X_{ij} = \begin{bmatrix} x_{1,1} & x_{1,2} & ... & x_{1,n} \\ x_{2,1} & x_{2,2} & ... & x_{2,n} \\ ... & ... & ... & ... \\ x_{m,1} & x_{2,2} & ... & x_{m,n} \end{bmatrix}$$

with $m$ variables and $n$ observations the similarity matrix $S_{ij}$ is defined as

$$S_{ij} = \begin{bmatrix} f(x_{1,1,i}, x_{1,1,j}) & f(x_{1,2,i}, x_{1,2,j}) & ... & f(x_{1,m,i}, x_{1,m,j}) \\ f(x_{2,1,i}, x_{2,1,j}) & f(x_{2,2,i}, x_{2,2,j}) & ... & f(x_{2,m,i}, x_{2,m,j}) \\ ... & ... & ... & ... \\ f(x_{1,m,i}, x_{1,m,j}) & f(x_{2,m,i}, x_{2,m,j}) & ... & f(x_{m,m,i}, x_{m,m,j}) \end{bmatrix}$$

where reflexivity of the similarity metric $f(x, y)$ implies that the diagonal of $S_{ij}$ is 1 and it's symmetry implies that all entries in the matrix are symmetric over the diagonal axis, assuming all entries in $X$ are non-zero.

To infer negative correlation, a dissimilarity metric is utilized for measuring opposing trends in observations between random variables. Here, a strong negative correlation represents the case where two variables are increasing and decreasing in value opposite of each other. A high negative correlation results from an increase in value for one variable indicates a reduction in the other variable. A low negative correlation means that if one variable increases there might be some degree of reduction in the other, but not substantial. The dissimilarity is close to zero when two variables become less dependent, and except from being defined in the number space [0,-1] it satisfies the same properties as the similarity metric [64].

Statistical correlation is a measure of dependence between two variables. It gives two pieces of information: The strength of the relationship and the direction of the relationship. The correlation value lies on a continuous scale between -1 and 1, any value close to either of these limits indicate strong correlation in any of the two directions. Correlation-based similarity measures are attractive candidates when dealing with large amounts of data because they are easily calculated and because the value can also be negative, it is able to distinguish between positive and negative relationships [65].

Comparison of high trough-put data by similarity measures to infer biological causality is not a straightforward task. Often there are a large number of quantified analytes and small sample sizes, so pairwise associations become spurious. This makes a true biological signal hard to identify. The probing of statistical associations found with various similarity measures needs to follow formal methods of causal inference [66]. High similarity of expression between two genes is formalized by a similarity metric, either by measuring linear or

non-linear dependence. To investigate interacting molecules, which are gene products, the gene expression profiles for a common set of genes between two compared conditions are compared - the *guilt-by-association heuristic* assumes two gene products share regulatory regime if they have similarities in expression profiles [67]. Similarity thus aims to identify co-regulation among genes because it has been shown to indicate functional similarity as well [68].

### 2.6.1 Spearman rank correlation coefficient

In co-expression studies, perhaps the most widely used similarity metric is the Pearson correlation coefficient which measures the degree of relationship between two linearly related variables [65]. A similarity measure that relates to the typical Pearson correlation coefficient is the Spearman rank correlation. This is calculated by replacing the observations by their *ranked values* and then computing the Pearson correlation coefficient of the ranks. Spearman's rho is given by

$$\rho_{ij} = \frac{cov(R(x_i), R(y_i))}{\sigma_R(x_i)\sigma_R(y_i)} \tag{2.12}$$

where $R(x_i)$ and $R(y_i)$ represent the ranks of variables $x_i$ and $y_i$, $cov$ denotes the covariance, and $\sigma$ is the standard deviation [69]. This is equivalent to calculating

$$\rho_{ij} = 1 - \frac{6\sum_{i=1}^{n}[R(x_i), R(y_i)]^2}{n(n^2-1)} \tag{2.13}$$

This metric does not carry any assumptions about the distribution of the variable observations due to the needed conversion of all observations for each variable into ranks, but is computationally slower than the Pearson correlation [70]. Nevertheless, because outlier values for a random variable will bias the correlation coefficient when this variable is compared to others, minimization of spurious outcomes is important. Spearman's rank correlation coefficient is less sensitive to outliers within data than the Pearson correlation coefficient. Spearman's correlation measures the extent of a monotonic relationship between two variables with less sensitivity to noise and occlusion [64].

### 2.6.2 Mutual information

Detection of genetic relationships based on mutual information (MI) can be done with estimation of the *entropy* between two genes. Computing mutual information between the RNA expression pattern for all gene pairs is done by estimation of the entropy of the expression pattern for each gene pair individually [71]. This produces a symmetric matrix of mutual information for all possible gene pairs in the data-set.

Mutual information is an information-theoretic measure that quantifies a statistical dependence between two random variables. Entropy measures how much information one can have about one variable, given information about the other. Given entropy knowledge of

one random variable, it quantifies the reduction in uncertainty of the other. In this thesis, MI is used to measure if knowledge about the RNA expression levels for one gene reduces uncertainty about the expression level of another gene. MI as a similarity measure was originally developed for discrete data. This type of data is characterized by having a *finite set* of possible states of which each state has a corresponding probability. For a random variable the information content for a feature is then calculated with the Shannon entropy (Equation 2.1). For a random variable $A$ with a finite set of possible states, $a_i, \ldots, a_{M_A}$, where $p(a_i)$ is the probability for the variable being at each of those states, the Shannon entropy is defined as

$$H(A) = -\sum_{i=1}^{M_A} p(a_i) \log(p(a_i)) \tag{2.14}$$

The joint entropy for $H(A, B)$ two random variables $A$ and $B$ is defined as

$$H(A, B) = -\sum_{i=1}^{M_A} \sum_{j=1}^{M_B} p(a_i, b_j) \log(p(a_i, b_j)) \tag{2.15}$$

where log is base 2 logarithm, and $p(a_i, b_j)$ is the probability A and B were within the state $a_i$ and $b_j$ of that feature, respectively.

Mutual information is then calculated for each pair, $M(A, B)$, according to the following definition:

$$M(A, B) = H(A) + H(B) - H(A, B) \tag{2.16}$$

The value for MI depends on the amount of information entropy within each gene's expression values. Genes who exhibit less entropy will thus have small values for M even if the two genes are highly correlated. High entropy means that the RNA expression levels for the given gene is more randomly distributed. Mutual information for a gene pair may be zero if and only if there is a strict independence between the two genes. It increases with less statistical independence between the two genes [72]. The relationship between two variables' entropy and mutual information is illustrated in Fig. 2.2 below, where the mutual information is termed $I(A; B)$.

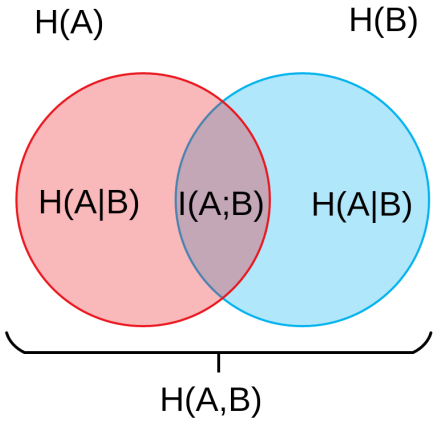


**Figure 2.2:** Diagram showing the mutual information and entropy. The figure shows the relationship between the two variables A and B and their entropies $H(A|B)$ and $H(B|A)$. The mutual information $I(A;B)$ is the sum of the individual entropies minus the joint [2].

Entropy is computed using discrete probabilities. But estimation of mutual information for random variables that are not discrete is not so simple. This is the case for gene expression data, which can take any value on a continuous scale given by the experimental setting and data processing programs determining any arbitrary unit of measured expression. Because the equations for estimation of entropy assumes knowledge about probabilities for the variables independently and the joint distribution (Eq. 2.1), for continuous data these unknown probability distributions have to be estimated.

Because mutual information is a measure based on entropy, it can largely be affected by small fluctuations in expression measurements because of biological noise. Because this similarity measure was developed for discrete data, continuous data is usually partitioned into discrete intervals. To ensure that estimations of mutual information is not confounded by biological or measurement randomness and noise, the continuous data is often divided into a large set of bins. Then the probabilities are estimated for each interval the expression data is binned to based on a computed relative frequency of data points in each of these bins;

$$\hat{p}(a_i) = \frac{1}{N} \sum_k \theta_i(x_u) \tag{2.17}$$

and the joint probability is calculated analogously with a two-dimensional histogram. This is the most straightforward approach. Because each data point is drawn from the data without replacement and assigned to a bin, noise in the data will affect which bin each data point is assigned to. Another limitation to the binning approach is referred to as the *finite size effect*, stating that MI estimated from a finite set will fluctuate around a higher value than the 'true' value [73]. This results in approximation logarithms of probabilities by logarithms of frequency ratios, and an estimation of $I(X,Y)$ by $I_{\text{binned}}(X,Y)$. This method is thus prone to be biased by the arbitrarily chosen borders of each bin and over-estimating the MI [74].

Instead of binning the data an alternative method to estimate the entropies can be applied

instead. This method is based on $k$-nearest neighbours statistics and is adaptive, being increasingly accurate for data sets of higher sample size. This method uses the average ranked distance to the $k$-nearest neighbour to estimate H(A), H(B) and H(A,B) separately as a method to infer the MI [75].

In contrast to conventional similarity measures, MI does not assume any linear relationships between the data. It provides a general measurement for dependencies in the data, making it very suited as similarity measure between genes that may all share responses but of opposing nature [76]. Some papers have shown that there is little difference between MI and the Pearson correlation measure [76, 77], but it has also been shown to outperform them in differentiation among clustering solutions [62].

### 2.6.3 Weighted topological overlap

The topological overlap of two nodes (genes) in a network is a measure of interconnectivity relative to the other nodes in network as a whole [78]. Equation for the weighted topological overlap measure:

$$w_{ij}^{TO} = \frac{\sum_k |a_{ik}a_{kj}| + |a_{ij}|}{min(k_i, k_j) + 1 - |a_{ij}|} \tag{2.18}$$

in which $k$ is the *connectivity* of a node and equals the sum of the weighted connections from a node to all its neighbors. The connectivity is defined as

$$k_{ij} = \sum_{j=1}^{n} |a_{ij}| \tag{2.19}$$

In the denominator of this equation, $1 - a_{ij}$ makes it mathematically unfeasible for the topological over matrix values to take negative values and in unweighted networks this value can only take the maximal value of $w_{ij} = 1$ if every neighbor of node i are all connected to node j, or node i is directly connected to node j. Conversely, $w_{ij} = 0$ is only possible if both nodes $i$ and $j$ are unlinked and they do not have any shared neighbors. To include the possibility of negatively correlated genes measured with wTO, the sign of the correlation value is multiplied with the wTO-value:

$$w_{ij}^{TO,F} = sign(a_{ij})w_{ij}^{TO} \tag{2.20}$$

The topological over matrix is, like the adjacency matrix, a symmetrical matrix in which all values along it's diagonal are zero, i.e. $w_{ij} = w_{ji}$ and $w_{ii} = 0$.

When weighted topological overlap is used as a similarity measure in gene co-expression studies, correlation measures for all pair-wise gene combinations are used to calculate the weighted topological overlap. It has recently been suggested to threshold the correlation

measures in order to arrive at relevant associations [79, 80]. This threshold is a number which becomes the exponent of each correlation value, makes each value more accentuated. Because a 'hard' threshold may lead to the loss of information [80] it is more common to apply the 'soft' thresholding proposed by Horvath [81]. This thresholding weighs each connection with the power adjacency function

$$a_{ij} = power(s_{ij}, \beta) \equiv s_{ij}^{\beta}. \tag{2.21}$$

Soft thresholding is performed with a power adjacency function, in which each correlation measure calculated with the ranked Spearman correlation coefficient is raised to the power of five. With $\beta = 5$, the signs of correlation measures are kept. This is done to retain negative correlations for the downstream identification of differentially co-expressed genes. Scale-free topology fitting index R squared and heterogeneity tend to increase with $\beta$ while density and centralization decrease with $\beta$ [81].

### 2.6.4 Overlap coefficient

The overlap coefficient or the *Szymkiewicz–Simpson coefficient* is a relative measure of overlap between two finite sets [79]. It quantifies the extent of similarity between set of data. For two sets of data $X$ and $Y$ the overlap coefficient is given by

$$overlap(X, Y) = \frac{|X \cap Y|}{min(|X|, |Y|)}. \tag{2.22}$$

The similarity measure is defined as $overlap(X, Y) \in [0, 1]$ [82]. In this thesis, the overlap coefficient will be utilized to facilitate assessments of compliance between compared results, see Chap. 5.1 and Chap. 5.4.2.

## 2.7 Gene co-expression networks

The motivation for gene co-expression analysis is its applicability to transcriptomic data for the identification of dynamic states of gene expression regulation in all kinds of cells. This analysis successfully describes relations between gene expression levels and functional states or cell behavior. The transcriptional regulation sets the rate of gene product production and essentially represents information processing in the cell. Analysis of this data can potentially identify predictive patterns of genetic activity: genes being expressed at similar levels in the same tissue, condition or developmental stage across a group of samples [11]. The genes highlighted as co-expressed across samples are likely to be functionally related. Their products may be physically interacting, or they can be implicated to take part in a specific process together. An important application of this analysis is the identification of interaction patterns related to cellular states, especially the system's transition from a functional state to a dysfunctional one [83]. These can be distinct cell stages and tissue

phenotypes perhaps associated with a disease. Expression data across dysfunctional states can thus shed light upon predictive genetic responses involved in orchestrating the molecular mechanisms a disease requires to propagate in the biological network. Applications of co-expression studies already at the micro-array period in the 1990s began to show that specific groups of samples could be described by their common sets of gene transcript levels [84].

The input to co-expression networks can be data from RNA sequencing analysis. Transcript fragments are sequenced and matched to the human genome, testing which genes are active and at which level they are being transcribed. Transcription data-sets presents gene expression levels across multiple samples per gene as a vector. Each element in the vector correspond to a read number; this is the abundance of a gene's transcript detected per sample. The complete set of samples, stored as elements in the vectors, is a collection of gene expression measurements in the same tissue. The mRNA abundances for one gene is thus contained in each vector, and the complete set of vectors for all genes constitute the entire set of data for an expression profile across the patient cohort [85].

Gene co-expression networks are employed to associate genes to biological processes in order to discriminate among genes according to relation types between them. In these networks the nodes represent genes and links between them represent *synchronization* in their expression. A co-expression network will enhance patterns of genes behaving in similar ways. System modeling of co-expression patterns for genes improves our understanding of this process. It a great tool to investigate intra-cellular molecular pathways, which essentially regulate all major cellular events, because it is designed to detect various co-occurring transcriptional profiles. These similarities in expression become descriptive of the data condition, e.g. the disease, from with the transcription network originates [85].

Construction of a gene co-expression network has two main parts: First, estimation of co-expression though a similarity measure $s(i, j)$. Similarity between expression levels for pairs of genes is measured with a statistical measure, e.g. Pearson's or Spearman's correlation coefficients. The similarity scores are stored in the similarity matrix $S = [s_{ij}]$. Inferred relations over a certain threshold for correlation are kept as indications of significant non-random properties of synchronized behaviour. As a convention the similarity matrix is converted into an adjacency matrix $A = [a_{ij}]$ in which all diagonal elements are set to zero. It is common to use the signum function implementing a 'hard threshold' $\tau$ to exclude relations with low significance. The signum function is given by [81]

$$a_{ij} = signum(s_{ij}, \tau) \equiv \begin{cases} 0 & s_{ij} \geq \tau \\ 1 & s_{ij} < \tau \end{cases} \tag{2.23}$$

Correlation is informative because it implies predictive behaviour of variable levels of gene expression. Second, these remaining associations form a list of gene-pairs used to find the of adjacency matrix. This matrix represents similarity scores for genes as rows and samples in columns [86]. Similarity matrices built upon correlations are symmetric and will this will

result in inference of undirected networks [84].

Co-expression analysis provides great benefits to research in molecular biology. In addition to integrating omics-data into a higher-order structure it can be used to elucidate *modular structures*. Modular network structures are neighborhoods of nodes that on average or connected to each other to a higher degree than to that of the rest of the network nodes. Co-expression networks have a great ability to group together genes that associate by various sorts of genetic interactions into modules. Some of these segregated gene groups may also be likely to share similarities in biological function. The network community provides a description of the local patterns of interactions among it's components. Modular structures have been shown to be evolutionary conserved and converging evolution of transcription networks of different species to contain modules of similar structure support the concept of biologically meaningful modular structures [87]. Local dense neighborhoods of co-expression networks are potential functional modules, thus analysis of these may lead to discoveries of novel molecular interactions and functional groups in biological networks [88].

There are several ways to perform analysis of modular structures in networks. Most are based on grouping nodes in communities based on a higher internal interconnectedness relative to that outside the potential community. Because there are potentially many possible ways of partitioning a network into smaller communities, there are various algorithms to perform this optimization problem. In this thesis networks are partitioned into communities by the Louvain-algorithm [4] implemented in the Python-package *networkx* [89].

The process in which normal cells develop into cancerous cells is exceedingly complex. Yet there are certain aspects characteristic of oncogenic processes that are key to differentiate normal versus carcinogenic cells. By studying the differential gene expression patterns of thyroid carcinoma, a deeper understanding of the genetic interactions and alterations in the cellular network can be reached [90]. Results from this analysis will potentially pinpoint individual genes and groups of genes that play important roles in the complex regulation of cellular function that differ between normal and carcinogenic tissues.

## 2.8 Differential gene co-expression networks

There is huge interest in comparison of RNA-sequencing data between various conditions. This type of study is termed *differential expression* analysis and aims to identify differentially expressed genes. Differentiated expression profiles are patterns of genetic transcriptional levels across many samples that have significant dissimilarities. Expression vectors containing measured levels of a gene for a set of samples may remain unchanged when compared to expression levels from the same gene measured in a different sample group. But for some genes there may be no similarity in expression between the sample groups at all. By comparing the expression profiles from two experimental conditions it becomes possible to elucidate which expression patterns are best characterizing the discordance. Differential co-expression identifies gene pairs with correlated expression profiles with specificity to a

given condition. When identifying these variances among expressed genes, it becomes interesting which role they have and what effect this difference has in relation to the two sample groups. They represent the two compared conditions. Some of the difference between the conditions can then be attributed to their particular expression profiles [11].

Construction of a differential gene co-expression network has the same initial steps as a gene co-expression network. Gene expression profiles are organized as vectors for which genetic pairwise co-expression is measured through a similarity measure, and from this data a selection of the most significant associations are drawn, yielding an adjacency matrix $A = [a_{ij}]$ for each condition. This part is performed analogously as when constructing gene co-expression networks described in the previous section. Typically, the two different data-sets analyzed represent a positive and negative in respect to a condition being investigated. They must contain expression vectors for an equal set of genes. Then they are comparable and can be employed to associate the characteristic expression patterns of specific genes to the condition under study.

Following there is an additional third step to achieve a quantitative comparison of expression profiles between the different sets of data. The co-expression scores from one transcriptomic data-set are compared to co-expression scores from different data-set. This step aims to identify the genetic sources of variation between the two sets of samples. This produces the similarity matrix of gene expression between the two data-sets. This is either done by creating a network from the adjacency matrix from each condition and comparing them, or by constructing a co-expression network from a new mutual adjacency matrix representing the *compared* property of the co-expression. In networks constructed by the latter method each value in the new adjacency matrix represents the change in co-expression pattern between the conditions. Hence, every pair of genes in the network sharing a connection has a link representing the differentiated co-expression.

Because gene expression must be highly specific to maintain the optimal energy use and functional requirements of the cell, the patterns of expression is indicative of function. For example, differential gene analysis between different tissues will highlight tissue-specific transcription profiles for some genes. These genes then will be especially important for the function of the tissue [91]. Tissue-specificity in differential co-expression networks also act as a natural filtering method to ensure that genes that are functionally more similar are included in the potential differentially expressed genes identified.

Analogously, differential gene expression between tissues representing specific phenotypes relate genetic expression patterns to phenotypic properties [92]. Many diseases have some specific gene expression abnormalities [93]. Employing differential gene co-expression on the study of diseases is very useful because it may identify significantly altered expression of genes specifically in the tissue the disease affects. The resulting network will thus be enriched in genes that mediate the sickness' network perturbation, and effectively point out the molecular characterization of the condition [84].

Lasty, a major benefit of differential gene co-expression analysis is to identify modular

structures within the network related to the onset of a specific phenotype - specifically in the application to study diseases. By studying communities within differential co-expression networks *disease modules* may be identified. These are statistically likely to be related to a disease if it's gene components are perturbed, for instance mutations and expression changes [88].

## 2.9 The CSD Method

The CSD framework has been developed by André Voigt and Eivind Almaas at the Department of Biotechnology and Food Science at NTNU [3]. The CSD method is a method for generating differential gene co-expression networks (DGCNs) highlighting internal distinctions of co-expression types. The method consists of a set of software programs for making an entire network representation of a DGCN discriminating among three different types of co-expression. In this network nodes represent genes and edges between them represent a conserved or changed value of co-expression for the given gene pair. A gene pair with high inter-sample correlation and low variance is more likely to be co-expressed. Edges thus reflect perturbations in expression patterns resulting from a condition-specific transcriptomic profile. By imposing expression patterns in this way, more information from co-expression measures can be used to recognize proper relations between conditions and detect expression profiles characteristic of a particular tissue state.

The analysis is based on assessing differences in pair-wise gene expression found between two data-sets representing expression profiles. First the similarity measure for every pair of the total set of $N$ genes present in the two expression files are calculated separately. The similarity measure holds quantitative information of similarity in expression levels for the gene pair $i, j$ with respect to a condition $k$. This measure is calculated with Spearman's rank correlation coefficient, $\rho_{ij,k}$. This measure can either show no correlation ($\rho_{ij,k}$ equals or is close to zero), positive correlation ($\rho_{ij,k}$ equals or is close to 1), or negative correlation ($\rho_{ij,k}$ equals or is close to -1).

The next step is to quantify the difference between the two conditions represented by the expression data cohorts. The CSD method compared to other co-expression analysis methods differs in that it incorporates information from co-expression measures between two conditions $j, k$ to prevail meaningful categories of differential expression patterns: the conserved ($C$), specific ($S$) and diverging ($D$) patterns of expression. Figure 2.3 shows a schematic explanation of the three types of co-expression categories. Discrimination between different types of co-expression refines the methodology in order to get a more comprehensive understanding of the mechanisms involved in the sickness. These types of expression relationships become the link attributes in the resulting differential co-expression network [3].

- *Conserved (C)* expression relationships represents situations in which a significant similarity of expression found in both data-sets. This means that the gene expressions correlation is found commonly in both conditions with the same sign.

- *Specific (S)* expression relationship represents a pair of genes for which only one and not both conditions is characterized by strong correlation of any type. Here, the correlation coefficient value for the gene pair is high for one conditions, whilst for the other condition the value is or is close to zero.
- *Differentiated (D)* expression relationship represents co-expression that differentiates for a gene pair across sample cohorts. In this case the correlation is strong in both conditions, but the sign of correlation is not equal.

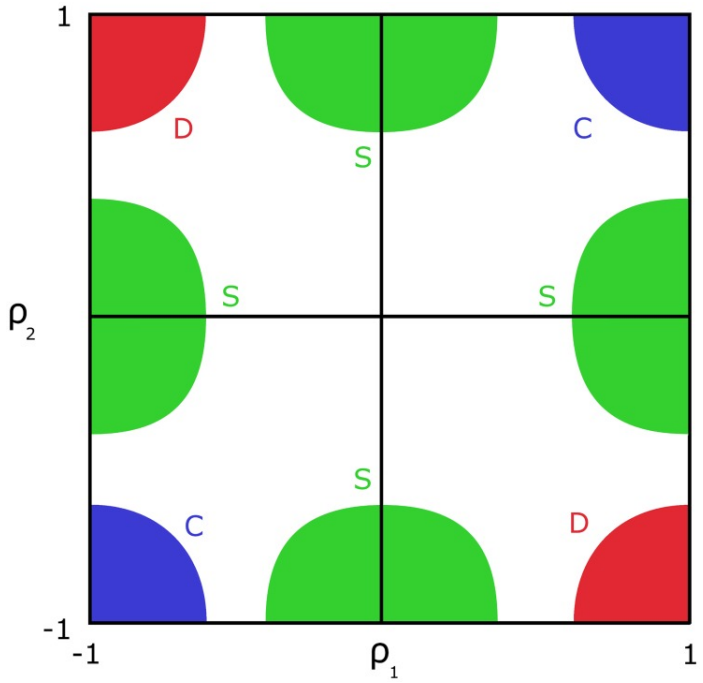


**Figure 2.3:** Diagram showing the regions of differential co-expression properties of gene-pair interactions inferred with the CSD method. For a given pair of genes their co-expression value measured by the Spearman correlation coefficient for condition 1 and condition 2 is denoted $\rho_1$ and $\rho_2$ respectively. The plot illustrates the different co-expression relationships between $\rho_1$ and $\rho_2$. These are categorized as either C-, S- or D-scores depending on the kind or relationship. C-scores (blue) describe *conserved* co-expression, with similar sign and both strong correlation. S-scores (green) describe *specific* cases where no relationship between co-expression values exist, with opposing strong values and opposite correlation signs. D-scores (red) describe *differentiated* cases where both co-expression values have strong values but different correlation signs [3].

This refined categorization of co-expression relationship between conditions makes the CSD method a powerful tool to identify condition-specific patterns. By imposing co-expression values between functional and dysfunctional states, it compares the transcriptional properties that causes it to be pathological and also highlights the types of relationships these transcription patterns have between the conditions. Edges of the network will reflect genetic interactions that have changed between a normal and impaired state, which in turn can be used to expose by which means the disorder manifests in the network and leads to the diseased state. The CSD method accommodates these node-relation changes in the network edges with the goal to highlight conditional changes resulting from a disease. The C, S, or

D relationship of co-expression is calculated by the following equations:

$$C_{i,j} = \frac{|\rho_{ij,1} + \rho_{ij,2}|}{\sqrt{\sigma^2_{ij,1} + \sigma^2_{ij,2}}} \tag{2.24}$$

$$S_{i,j} = \frac{|\rho_{ij,1}| - |\rho_{ij,2}|}{\sqrt{\sigma^2_{ij,1} + \sigma^2_{ij,2}}} \tag{2.25}$$

$$D_{i,j} = \frac{|\rho_{ij,1}| + |\rho_{ij,2}| - |\rho_{ij,1} + \rho_{ij,2}|}{\sqrt{\sigma^2_{ij,1} + \sigma^2_{ij,2}}} \tag{2.26}$$

The values for each of these scores, which are quotients of expressions with absolute correlations as numerator and the root of the summed expression vector variance of the gene pair's as denominator, are anywhere in the range from zero to infinity. Therefore the distribution of scores included in the final network construction with all three co-expression types are included based on a cut-off threshold computed for each type respectively. Because the C, S and D-scores follow different distributions, they must be converted into comparable measures that can be integrated in the same network. A common importance level $p$ for each score is needed, with corresponding threshold values for each type $k_p^{C,S,D}$. Values from each collection of scores are discarded if they are below this threshold. The cut-off values should be computed in a way so that none of the scores have overlapping areas in respect to each other and because this threshold necessarily affects the number of edges and nodes in the resulting network construction the cut-off values should be adjustable according to potential requirements of downstream network analysis.

The thresholds $k_p^{C,S,D}$ were calculated in the following way: All elements within each distribution are all combinations of gene pairs $M$ from the total set of $N$ genes. From the $M$ data points, a sample $s_i$ is drawn $m$ times. Each set has a size $L$ where $L << M$. This results in a set of $M$ different values for all three $C$-, $S$- or $D$-scores. The threshold value is then calculated by averaging the maximal values per sample for each type. As example, threshold $k_p^C$ is calculated as: $k_p^C = \frac{i}{m} \sum_{i=1}^{m} max_{s_i} C$.

An additional strategy for computing the equations 2.22 - 2.24 will be explored in this thesis. This alternative method for finding the $C$-, $S$- and $D$ association type scores of genetic relationships is given by the same set of equations, but with all denominators equal to 1. This is equal to computing the three scores without correcting for independent variance of each gene's expression values for the $n \times n$ gene pairs. Because the estimation of gene expression vector variance for data-sets consisting of typically $n = 20,000$ genes must be done individually with a proper sub-sampling algorithm (described in 2.9.1), this step clearly poses computational demands. Hence, skipping this step could be advantageous if not impeding the quality of the genetic relationship inference strategy. This alternative will be referred to as "CSD-VAR" because it only differs from the original network inference algorithm by the variation correction. The effect of correction for expression vector variance will be investigated by comparing results from applying both the original inference algorithm

and this alternative one to the same expression data-set.

### 2.9.1 Sub-sampling algorithm for variance estimation

The similarity in gene expression levels measured with the Spearman correlation yields a $\rho_{ij,k}$-value for each gene pair $ij$ for condition $k$ for all data points $N$. Each of the scores $C_{i,j}$, $S_{i,j}$ and $D_{i,j}$ are measures of changes in $\rho_{ij,k}$ between the healthy and sick tissue and must therefore be corrected for variance within the gene expression values for each gene pair. Confounding factors may affect the variability of $\rho_{ij,k}$. It must therefore be determined and corrected for. Each of the scores are divided by the sum of the variance from the two conditions. The scores that have high internal variance for one condition independently will then reduce the value of the computed C-, S-, or D-score and be filtered out later in the process by a hard threshold. In this way, experimental or batch effects altering the variability within the expression values for one gene pair will not swamp the expression value variability for other gene pairs. If the correlation was not corrected for variance within the set, there sould be an increased probability of assigning co-expression correlations between the conditions for a gene pair due to other factors than the conditions. The correction for variance in gene expression values will thus remove bias from the correlation measure, so that observed patterns in co-expression will be more likely to be caused by the difference between the experimental conditions; thyroid cancer and normal controls.

There is a dependency between the correlations measure and a sub-sample variation. If the absolute value of correlation for a gene pairs expression values is high, the variation within the sub-samples should be smaller. A biological explanation for this is that there should be a degree of consistency in regulation of genes that indeed share biological function; by being regulated by some common transcription factor or that having gene product of similar function will be reflected in a high correlation value. This also logically implies that a low variability between the same genes can be expected, rendering the similarity score for this gene pair high relatively to others.

To estimate the variance in co-expression for each variable in the correlation measure, the correlation measure is computed for a set of sub-samples of size $n$ randomly drawn from the set of $N$ data points. If there is a large internal variance in gene expression values for a gene pair, the corresponding $C_{i,j}$-, $S_{i,j}$- and $D_{i,j}$-scores should be reduced accordingly. The variance is computed with the square of the standard deviation of the mean for each sub-sample. The sub-sampling approach needs to to maximize the number of achieved sub-samples drawn and maintain independence between each of these sets. The following describes the implementation of the sub-sample selection process:

1. The total set of expression data points for each gene are ordered and sequentially numbered.

2. Each set for each gene is divided into sub-samples of size $n$ so that each value is contained in one sub-sample only.

3. The first data point in each sub-sample is used as initiating data point, $n^*$, for building a new sub-sample. Iterating through the data points, new points are added only if they have not co-occurred in a sub-sample with any of the other data points in this new set.

4. Step 3 is repeated so that each new set contains a unique combinations of data points of size $n$.

5. When there exist no new unique combination of data points initiation with $n^*$ to create a sub-sample, step 3 is repeated using $n^* = n^* + 1$ as the new initiating data point.

6. The process is finished when there exist no more valid sub-samples of size $n$ that can be created and $n^* = N$.

To ensure meaningful calculation of Spearman correlation, a large number of sub-samples with as many data points in each sample should be selected. In addition, because the Spearman correlation loses accuracy for smaller sub-sample sizes, there should also be consistency between the chosen sub-sample size for each experimental condition.

### 2.9.2 Node homogeneity

Any node in a network can be described by it's composition of associations to its neighbor nodes of different types. In a co-expression network constructed with the CSD framework, the three different link types are $C$, $S$ and $D$-links. These make up the three possible categories of types in the link type distribution attributed to a node $i$. This distribution is termed the *node homogeneity*, $H_i$, and is given by:

$$H_i = \sum_{t \in (C,S,D)} (\frac{k_{t,i}}{k_i})^2 \tag{2.27}$$

where $k_{t,1}$ denote the number of links of type $t \in (C, S, D)$ respectively and $k_i$ is the node degree. The homogeneity describes the node's interaction type properties. It quantifies whether a node (gene) for which co-expression correlations are predominantly of one single type, or rather with close to equal amounts of links of types $t \in (C, S, D)$.

The first case represents a *homogeneous* gene, which will associate with other genes in only one way for both of the studied conditions. An *non-homogeneous* gene will have a set of links of heterogeneous types. The links will then be of more than one type and none of the types relatively dominant compared to the others. This gene's expression associates with other gene's patterns in a variety of ways depending on the neighbor gene.

## 2.10 Network medicine

For many diseases the cause can be mutations or smaller changes in the DNA. The genetic material is a very specific sequence, 3 billion nucleotides long, distributed on the 42 molecules of nucleic acid that lies within the nucleus of every cell in the human body. These nucleotides' special sequence encode genes; these are the cell's instructions for synthesizing functional molecules, and they are the result of millions of years of evolution. Alterations disrupt genes and limit the cell's ability to behave properly. Cells then express proteins that behave abnormally, driving the development of the diseased phenotype [91]. For example, a mutation in a any one of 50 genes encoding retinal proteins may cause the disease retinal pigmentosa [94]. The damaged gene is translated into a protein that doesn't function properly and ultimately lead to vision loss. The genes that encode proteins that are needed to build the retina are especially active in the cells of the retinal tissue. Regulation of genetic transcription is essential for normal cell function. This ensures that in a normal retina, cells are producing photo receptors for the retina and not - for instance - collagen for the growth of hair-fibres. It is generally acknowledged that specific tissues have increased expression of the genes encoding proteins that carry out reactions essential for the functionality of the given tissue [12].

One might then assume that these altered tissue-specific genes are expressed and damaging exclusively their functional tissues. But this does not always hold true, and complicate the simple dogma of gene-cause-disease associations [92]. Any inherited mutation or genetic variation associated with disease exists in every cell of the human body, not just in the tissue in which the disease is typically manifested. Many genes linked to diseases are in fact expressed in multiple tissues throughout the human body, most of which not manifesting the functional abnormalities linked to the disease. This must mean that either the genes does have malignant effects in other tissues that are either masked or hindered in making harmful outcomes, or the gene cannot mediate the diseased phenotype on its own. If the gene cannot single-handedly make a tissue abnormal itself, it must depend on other factors to manifest the sickness. In other words, the expression of a gene associated to a disease does not explain the full mechanism of pathology in a tissue: The gene is not exclusively expressed in this tissue yet only portray the disease here [91].

This is why biological network analysis becomes a powerful tool in investigating human diseases. Analysis of disease-specific networks thus makes it possible to elucidate the characteristics of the genes that do manifest disease in specific tissues and the mechanisms by which it plays out its tissue-specific role. Comparison of genetic networks of normal and sick tissue will show the differences between the functional paths in which the genes interact. It will potentially also show why some genes cause pathophysiological changes in a single tissue while other affect multiple tissues. The goal is to identify gene network sub-clusters that bundle specifically in specific regions and depend on the integral expression of other gene members of the sub-graph relevant for the molecular mechanism of disease manifestation. The partners in a cluster like this will depend on the tissue. Also, the effects in tissues other

than the tissues in which the disease-associated genes usually are expressed most can be analyzed similarly; its internal structures compared to the disease-specific tissue. This will show why other tissues are less - or not at all - affected by genes associated with disease.

Network analysis of differential co-expression networks representing diseases have also been used to identify particularly interesting modular structures. These networks have demonstrated that genes encoding proteins related by distinct types of similarity tend to interact with each other. These may be found grouped together in network clusters [92]. Network clusters like these can be enriched for genes encoding products with similar functions, called *functional modules*. Recently, an even more interesting type of cluster has been hypothesized as well.

### 2.10.1 Disease modules

A *disease module* is the manifestation of a disease in a network mechanistically linked to a particular disease phenotype [88]. In a differential co-expression network it is a sub-network structure containing multiple disease-associated genetic entities. A great application of this theory is that a gene already linked to a disease can be used to identify novel disease-associations for other genes if they are found in this first gene's network neighborhood [95]. Analysis of these modules and their engagement in the network may be employed to identify the mechanisms carrying out the pathological phenotype. Also, in tissue-specific network they can potentially expose why certain genes manifest diseased phenotypes in some tissues but not in others [92].

To identify disease-modules within a differential gene-expression network, there are some criteria set for the network analysis: First of all, the network represents the genetic interactions in a tissue the disease is manifested. Disease-associated genes are highly connected and localized in the same neighborhood of the genetic network, forming this disease module. When expressed in other tissues, these genes have other interactions and are segregated into other regions or the genetic network. In a network representation of the tissue the disease typically affects the most, the module's integrity determines the cellular outlook characteristic for the given disease [88].

## 2.11 Statistics

### 2.11.1 Hypothesis testing

A hypothesis formulating an assertion is tested with the procedure *hypothesis testing*, in which two potential hypothesis are tested against each other. A hypothesis could be that a parameter of a population has a certain value, forming the alternative hypothesis $H_1$. The null hypothesis $H_0$ would then be the presumed value for the parameter. To determine which hypothesis is true, computing a critical value of $H_0$ is used to deduce rejection and

acceptance regions. These define limits to the test statistic. If it is in the reject region, $H_0$ is rejected for the alternative hypothesis $H_1$ [96].

For example, when comparing gene expression patterns looking for correlated behaviour, the alternative hypothesis is that two genes are correlated and share a linear relationship. Testing this hypothesis is done by comparing their correlation coefficient with that of a strictly uncorrelated patterns with similar parameters. If the difference is within the acceptance region for $H_0$ it is kept and the alternative hypothesis is rejected. But if the difference between the correlation coefficient for the two genes is located in the reject region for $H_0$ this indicates that their patterns are significantly more similar than entirely uncorrelated and the alternative hypothesis $H_1$ is accepted on the basis of the observed relationship between the gene expression profiles.

Hypothesis testing needs a significance level $\alpha$ to find thresholds for accepting or rejecting the null hypothesis. The significance level represents the rejection region of the null hypothesis. The $p$-value in this example represents the probability of observing correlations between actually non-correlated patterns, that is observing significant values of correlation by chance. A p-value below 0.05 is considered to be significant, because it is lower than the significance level. An even lower p-value is a highly significant indication of true correlation. If the p-value is lower than 0.05, more than 95% of the time correlation coefficients between these genes would result from significantly correlated gene expression patterns. The null hypothesis can then be rejected in favor of the alternative one, because the p-value is lower than the significance level $\alpha = 0.05$.

When carrying out hypothesis testing the probability of rejecting a null hypothesis that is true is equal to $\alpha$. A test resulting in a rejected true null hypothesis is termed a type I error, or a false positive result. If the alternative hypothesis is true but rejected it is called a type II error, or a false negative result. Following the same example, a false positive result when performing gene co-expression analysis would be to infer relationships between genes that in reality share no common biological behaviour in terms of transcription or regulation. A false negative represents the case where no association is inferred between genes that actually have correlated transcriptional behaviour.

### 2.11.2 Testing and correcting for multiple comparisons

When comparing large gene-expression data sets for pairwise correlations it becomes a huge set of simultaneous hypothesis tests, referred to as *multiple comparisons*. The probability of type I error increases with the number of comparisons tested at the same time. Thus, is becomes important to perform proper correction methods to account for this effect [97].

One common way of controlling type I error is with the *false discovery rate* (FDR) which measures the expected ratio of incorrectly rejected null hypothesis among all rejected hypothesis done during multiple comparisons. When performing comparisons of multiple gene expression profiles testing for significant correlations, accepting a maximum of 5% false pos-

itive assigned relationships between genes the FDR would be set to 5%. There are different strategies to control for the FDR. Here, the Benjamini-Hochberg method and the Bonferroni method will be described briefly [97].

The Benjamini-Hochberg method is named after Benjamini and Hochberg who developed this technique in 1995 [98]. Individual p-values are ranked in ascending order and compared to the critical value

$$\frac{i}{m} \times Q$$

where $i$ is the rank, $m$ is the number of tests, and $Q$ is the FDR, .e.g. 0.05. All test with p-values smaller than and up to $p < (1/m)Q$ will be considered significant.

Another method of correcting for error rates is the Bonferroni correction. This approach controls the FDR by the error rate for the complete set of tests, referred to as the *familywise error rate*. The critical significance value becomes lower than $\alpha$. Then if for all the multiple tests the null hypothesis $H_0$ holds true given this lower $\alpha$, the probability of having at least one false positive in the family of tests due to chance is 0.05. The Bonferroni correction is done by dividing the critical value by the number of tests. For example, for a set of 100 tests for which $\alpha = 0.05$, the new critical value becomes $0.05/100 = 0.0005$, being the new p-value threshold for accepting a test as significant [99].

# Chapter 3

# Methodology

## 3.1 CSD Analysis on Thyroid Cancer

The first aim of the research topic for this thesis was application of differential gene co-expression analysis of thyroid carcinoma. This section will describe the data pre-processing and implementation of the CSD method to infer differential co-expression networks. The strategies for extracting information from these networks in the aspect of disease investigation will be described in detail.

### 3.1.1 Data set collection

The gene expression data for healthy tissue used in the analyses of this thesis was downloaded from NIH Genotype-Tissue Expression (GTEx) Consortium V7, in fully processed and filtered single-tissue gene expression matrices in .bed-format [100]. The Genotype-Tissue Expression (GTEx) program is a multi-center effort to generate genomic and transcriptomic profiling data for more than 50 tissue sites from hundreds of autopsies [100]. The samples were collected from healthy human patients' thyroid gland tissue. The data set consisted of a series of 399 samples of gene expression data based on the GENCODE 19 transcript model. Gene-level quantification of expression in transcript per million (TPM) values were produced with RNA-SeQC v1.1.8 [101] for gene-level expression quantification. Transcript-level quantification was calculated with RSEM v1.2.22 [102].

The gene expression data of thyroid carcinoma was collected from the Thyroid Cancer project (THCA) from The Cancer Genome Atlas [103]. Expression data from thyroid primary tumor tissue diagnosed with various types of thyroid carcinoma were identified and accessed through the GDC Data Portal. In total 504 HTSeq-count expression files were downloaded.

In the THCA project there was also expression data available from solid normal tissue collected from the same patients with thyroid cancer, but these 70 data samples were excluded

from the analysis. A total of 504 expression files in HTSeq-format were downloaded and assembled into a large data set with gene expression data for 60,483 mapped reads.

The data set used in this analysis had the following composition: 355 patients were diagnosed with Papillary adenocarcinoma. 102 patients were diagnosed with Papillary thyroid carcinoma, 38 of which with the follicular variant and 9 with the columnar cell variant. 135 samples of the data set were from male patients and 352 were from female patients.

### 3.1.2 Data set integration

Because the data from TCGA was divided in separate files for each patient, the data was collapsed into one complete data set for all patients in the cohort. On TCGA the expression data from the THCA-project contained transcriptomic data from patients diagnosed with thyroid carcinoma, but not all of the samples originated from carcinogenic thyroid tissue. From the clinical data downloaded with the expression-files, the 66 transcriptome files in benign tissue (from the same thyroid cancer patients) were identified. This clinical data-file was used as a template for reading in the correct transcriptomic files, so that expression files for benign tissue were omitted, and the remaining 504 files from carcinogenic thyroid tissue was included in this study. The transcriptome files of solid normal tissue that were available on TCGA originated from patients in the THCA-cohort, these were matched tissue samples from the same patients but in healthy tissue. Exclusion of these files was based on the intention of keeping the cancer patients and healthy control patients statistically independent. In this way, there are no shared patients in the two compared groups of the study and there is virtually complete independence between them.

For differential co-expression analysis the two data sets should ideally consist of the same set of identifiers for the same genes and a similar set of genes. One challenge when using data sets from different databases is that there may be ID conflict between the sets and differences in data set compositions. Both the normal control data from GTEX and the THCA-data from TCGA had Ensemble identifiers. But, in order for the results of the differential gene co-expression analysis to be easy to interpret, all gene IDs were converted to the official gene symbols.

The script for this conversion was written in Python, converting all Ensemble gene IDs into the official gene symbol, based on the GENCODE annotation version 38 [104]. Even though there are some ID mapping services available online, these did not handle isomers in the desired way. In the expression data more than one instance of the same gene was represented because the same gene may have several different transcripts depending on splicing events and these mapped to different probes accordingly.

To address this, both data from THCA-patients and healthy controls, gene expression data from multiple probes were collapsed into a single unique value for the common gene name. Transcription isoforms originating from the same gene but with various lengths were converted into one gene expression value for all probes. This was done by calculating an average

value per patient for the same gene and this average was kept. This averaged the expression levels for transcripts of different lengths. After this was done, all expression vectors per gene had an equal amount of elements as the number of patients in each of the cohorts, 504 for the THCA-patients and 399 for the healthy controls. This process was chosen both because of easy data set quality and version control later in the process, but also because a differential co-expression analysis at single-gene resolution was sufficient.

Then the two data sets were filtered to contain a common set of genes, and for the downstream analysis the expression value vectors for the common set of 24,753 genes were included in the transcriptome data files for both thyroid cancer patients and the healthy controls.

### 3.1.3 Pre-processing procedure

The general agenda of the pre-processing was to remove batch effects across RNA seq data data from different sources. This quality control wok-flow consisted of multiple steps. All Ensemble IDs were converted to official gene symbols. Gene expression data for the thyroid cancer patients was normalized and filtered the sick tissue in as recommended in [105]. Expression values originally in raw reads format were filtered to include only genes with expression values $> 6$ reads in at least 20 % of samples. These reads were scaled to upper 75%-quantile library scale resulting in transcript per million-units (TPM). Only values over 0.1 TPM in at least 20% of samples were kept. This thresholded for a change in expression level set to $log|foldchange| \geq 2$. TPM values were corrected for gene length and converted to TMM with edgeR [50]. Lastly, the expression levels for each gene were transformed across samples with an inverse normal transform. The R-package edgeR was used because it implements possible bias sources into the statistical model to perform an integrated normalization [43].

As last part of the pre-processing, the two data sets from normal and THCA tissues were filtered to contain the same set of genes, and thus the common set consisted of 24753 genes in each set. The healthy GTEX data with 399 samples and the 504 samples THCA data from TCGA were used in all analysis of this thesis.

### 3.1.4 Differential co-expression analysis workflow

The CSD framework developed by Voigt and Almaas [3] is written i C++. The code computes pairwise correlation values for all pairs of genes and estimates variance for each gene's expression vector. The latest version of this code converted all values into ranks as a first step and then computed correlations for these ranked values instead of converting one by one expression value into a ranked value before computing the correlation to other genes. This is a faster implementation than the original, which is available on GitHub [106]. Spearman correlation coefficient is the similarity applied to calculate correlation between

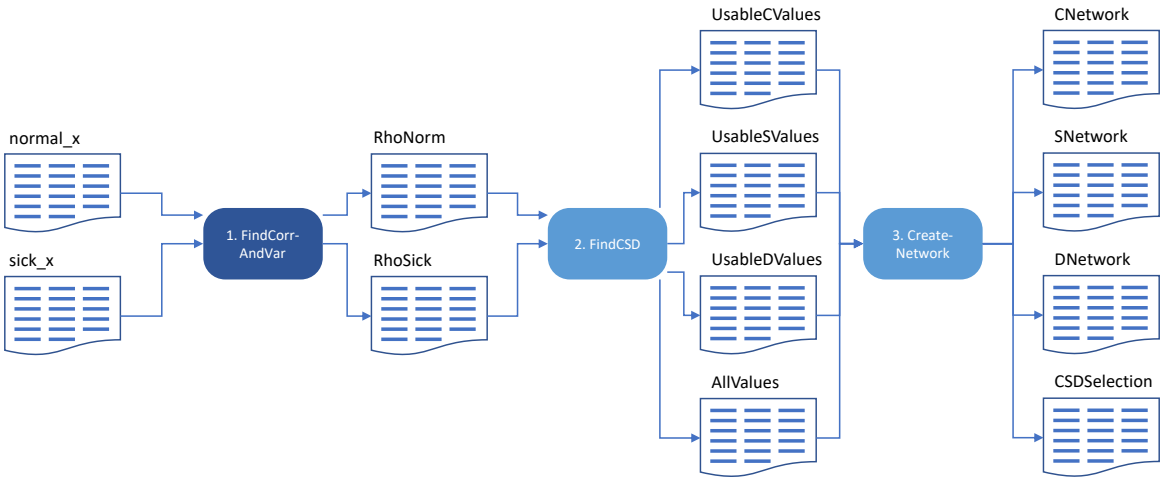


**Figure 3.1:** Illustration of work-flow for differential gene co-expression analysis with the CSD framework developed in [3].

differentially expressed genes between the healthy patients' thyroid tissue and malignant thyroid gland tissue.

The expression data from patients with thyroid cancer and the healthy thyroid tissue expression data was used to generate differential gene co-expression networks. The pairwise Spearman correlation coefficients were computed for each data set independently as a first step. This was done by compiling the C++ code and filling in the correct name of the input-file, specifying how many genes and samples the data set consisted of. A third parameter that was altered was the number of samples per sub-sample used to calculate the variance. This was set to 8 if the sample size was higher than 50. This code utilizes the sub-sampling algorithm described in the theoretic background to estimate variance, which should have at least 7 samples in each sub-sample group to estimate achieve three-digit accuracy of the Spearman correlation [3]. This is step 1 in Fig. 3.1, called "FindCorrAndVar".

The output from this first part generated co-expression for each separate condition. These were files with each gene pair and it's correlation value and variance. The next part was to use a Python script to compare these two correlation files from the two conditions and calculate all C-, S- and D-values. This refers to the Python-script "FindCSD" in step 2 in Fig. 3.1. This was done by specifying the names of the out-put from the previous step as input and reading in one line per gene-pair $ij$, corresponding $\rho_{ij}$-value and $\sigma^2_{ij}$-value per condition and calculating these three scores based on this information.

As a third step these C, S, and D-values are used as input as the last Python script which generates four networks, each for the three interaction-types and one aggregate network containing all the interactions of all three types. This is the Python-script "CreateNetwork" in step 3 in Fig. 3.1. The interactions included in these four networks are exclusively those who are above the computed threshold importance level for each interaction type, $k_p^{C,S,D}$. This code uses a parameter called *selSize* which is equal to 1/(desired p-value). The threshold is set to average maximum of the C-, S- and D-scores individually, drawn from a random subsample of size $1/p$ where $p$ is the importance level. Elements from the complete

set are selected with replacement [3]. Because each data set has approximately 20,000 genes with around 400 or 500 samples, the entire set of gene pairs is $\frac{n*n-1}{2}$ and the sub-sample sets were of size $1/p$. Thus, given a importance level of $10^{-5}$ this is around 0.001% of the total set of gene pairs. From healthy and thyroid cancer co-expression data differential co-expression networks were inferred on the basis of four different values for *selSize* yielding networks based on importance values $10^{-4}, 10^{-4.5}$, $10^{-5}$, and $10^{-5.5}$. Resulting networks were visualized in Cytoscape [107] for inspection and analysis.

### 3.1.5 Box-plots

Box-plot are graphical representations of data in which multiple relevant properties for each sample are included in the visualization. It is a graphical tool which characterises the distribution of the numerical data with percentiles, means, medians, and minimal/maximal values. A box in the plot represents a sample, the length of this box is the inter-quartile range covering the central 50% from the first quartile to the third, while whiskers may be extended out from each box to the minimal and maximal of the values. On a vertical box the mean, given by the second quartile, is indicated with a horizontal line inside it. The median of the data and any outlier points may be also be indicated in the plot [108].

### 3.1.6 Node homogeneity analysis

The homogeneity in the complete network was investigated by calculating the homogeneity score for each of the genes in the network with Eq. 2.27. The homogeneity for each node depended on its distribution of links of the three different types; conserved, specific or differentiated. For all genes in the network the homogeneity was calculated and then these values were sorted by the node degree. Then the homogeneity was plotted as a function of node degree. The result from this analysis is shown in Fig. 4.4. Here all homogeneity-scores are shown in box plots. For each node degree the spread of the collection of homogeneity-scores determines the length of the boxes, its length along the vertical axis represent the first ad the third quartiles of homogeneity-scores. The maximum and minimum values are indicated by the whiskers extending out from the top and bottom of the boxes respectively. For the homogeneity-scores per node degree the median and the mean values were computed, these are also included in the box plot. Red horizontal lines denote the median and the green bars the mean.

### 3.1.7 Gene ontology enrichment analysis

Gene ontology (GO) biological process enrichment analysis was performed for the complete differential gene co-expression network inferred with the CSD method. This was done to assay the differentially co-expressed genes (DEGs) between thyroid cancer and healthy persons for enrichment of biological processes. In the network the genes were denoted by

their official gene names. Converting these back to Ensemble IDs resulted in a higher ability of the GO enrichment tools to map the gene names to their genetic databases, and was therefore done for all sets of genes before performing the GO enrichment analysis. The enrichment analysis was performed using the tool DAVID [109] and the GO Enrichment Analysis powered by PANTHER [110].

The results from biological enrichment analysis provide lists of biological processes with significant enrichment for the set of differentially expression genes it is given. The test asks if there among the DEGs are more members of any biological or functional pathway than what would be expected in a uniform random selection of genes. It is thus a statistical approach to identify over-representation taking differential expression into account [68]. The list usually provides test significance level, false discovery rates, and the Bonferroni correlation coefficients for each identified biological process. GO enrichment analysis is a method of testing which biological function a group of genes are involved in. The analysis will use the protein-coding genes from the queried gene list to find the pathways these genes' products take part in and thus map the characteristic behaviour for the group of genes in the organism [68].

### 3.1.8 Community detection

Investigation of module functionality and potential disease associations was performed by network community detection of the network constructed based on CSD. This was done by implementing the Python *communities* package which employs the Louvain-algorithm to find communities [4]. Community detection was done by using the package's function "best partition". Each module was written out to a file and from this file the module affiliation for each gene symbol was stored as attributes. Lastly the nodes in the modules were coloured by their module affiliation in order to investigate module partitioning of the network. Results are presented in Fig. 4.7. Network community detection is a tool that aim to identify functional modules, which may potentially aid in the identification of disease modules. Resulting modules were examined for presence of disease-associated genes.

### 3.1.9 Detection of disease genes and potential disease modules

The functional annotation tool PANTHER [110] was used to identify genes in the differential gene co-expression network associated with thyroid carcinoma. These were recognized as DEGs based on the CSD method for network inference with importance level $p = 10^{-5}$. Of the 7224 DEGs from the CSD network of Analysis 2, 1308 Ensemble IDs were successfully matched with the Gene Association Database [111]. Of these 22 genes were identified as thyroid cancer-associated with a p-value $= 8.4 * 10^{-5}$ and Benjamini corrected p-value $=$ 0.065.

Examination of genes associated to thyroid cancer was an integral part of the general aim of this thesis. The investigation of them could potentially elucidate novel knowledge about

thyroid carcinoma. In addition, the relative quantities of these disease genes in network neighborhoods could identify cliques with a communal association to thyroid carcinoma as well. These network neighborhoods could be strong indicators of functional - hence a *disease module*. To this end, network clusters were examined for presence of these genes. Lastly, the ability to identify such genes could act as a robust way of quantitative assessment of different similarity measures as basis for differential gene co-expression analysis.

## 3.2 Method study and development

The second main objective of this thesis was to study and expand the methodologies for performing differential gene co-expression analysis. First, the effect of proper pre-processing was examined. Second, new implementations of different similarity measures were developed. In this method development part, novel similarity measures were introduced in a way maintaining most of the originally CSD framework. By not changing too many constituents of the work-flow, quality control is more easily feasible and comparison of results generated with these alternative strategies will have higher validity. Faster implementation of weighted topological overlap was attempted and the mutual information was developed as a new similarity measure as foundation for the CSD framework. Also, a simpler version of Spearman's correlation coefficient was examined. The effect on the differential gene co-expression analysis outcomes was explored for each similarity measure.

The transcriptomic data used in this section was identical to that collected in Chap. 3.1.1 and integrated into a data set as described in Chap. 3.1.2.

### 3.2.1 Alternative methods for data pre-processing

Pre-processing methods have an important impact on the quality of the data provided as input for co-expression analysis. Various methods for modeling of transcriptomic data and pre-processing for differential expression analysis address statistical challenges differently. An illustration of these pre-processing strategies in provided in Figure 3.2. The initial step of the pre-processing process common to all alternatives was to convert Ensemble IDs to official gene symbols, see step 1 in Fig. 3.2. After this, three somewhat different filtering procedures were performed as part of quality control of the expression data, resulting in three distinct data sets of thyroid cancer transcription data. All of these were both used in the CSD workflow after pre-processing. Here blue boxes represent processes implemented in Python and green boxes represent processes implemented in R. The two filtering processes differed by filtering stringency. All processes were to a certain degree replications of the normalization method already applied on the healthy control transcriptomic data downloaded from GTEX.

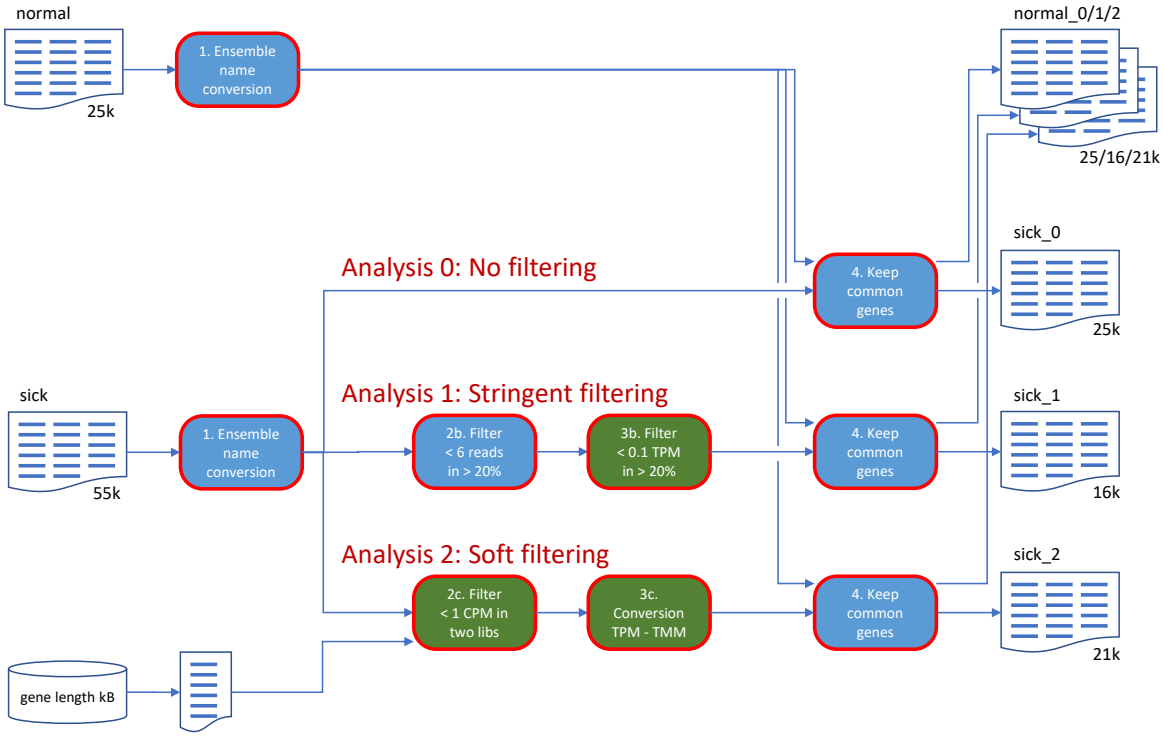


**Figure 3.2:** Illustration of alternative pre-processing steps.

The first step of the of cancer pre-processing was to filter out low read-count genes from the data, see step 2b in Fig. 3.2. All read count values per gene that had less than 6 raw read counts for more than 20 % of the samples were identified. This corresponded to genes whose expression value vector had more than 100 instances of the value zero. These were removed from the data set. The next step was to normalize the transcriptome data. Normalization of the RNA-seq raw count values was done by further implementing the same method with which the healthy RNA-seq data from GTEX had been normalized according to their expression preparation procedure [112]. This included both conversion of the raw gene counts into reads per kilobase, then transcript per million values and normalization using the R package edgeR [50]. The workflow will be described in detail in the following section.

First, conversion of the raw gene counts from RNA-seq data was done by conversion to read per kilobase-values (RPK) an scaling into transcript per million (TPM). Gene lengths for all genes were obtained by querying sets of genes on GeneALaCart [113] or computing the length from GENCODE annotation files, which included start and end position on the chromosome, se bottom process in Fig. 3.2. Gene lengths in kilobases for each gene were calculated and stored in a dictionaries for each of these sources. Because of various transcription lengths from the same gene, each source held multiple gene length values for each gene which were all stored in a list. But in order to convert transcriptomic data into RPK values, one single gene length was needed. Thus, the best alternative gene length from all transcription versions had to be found. For a set of 100 random genes, their lengths were assayed manually against the UniGene NCBI database to find which value was closest to the reference length. From this it was found that the best estimate of true gene length was

obtained by keeping the maximal value of gene length in case multiple values were found per gene. Then these gene length-dictionaries were merged to hold only one set of unique genes with one maximal gene length value for each of them. The total gene length in this collection was 62,910. All transcriptomic data in the thyroid cancer data set was divided by it's per gene length in kilobases, to yield RPK-values.

Second, all RPK values per sample were counted and divided by one million, which produced a set of per million-scaling factors. Third, all RPK values were divided by the per million-scaling factors to yield expression values in TPM units. As a final step, the values were also transformed with a inverse normal transform function. These last steps were performed in R. As mentioned above, this conversion process was done exclusively for the thyroid cancer data, because this was in raw read count units when downloaded.

The two data sets used in the differential co-expression analysis were different in that the normal data from GTEx was pre-processed and the data from TCGA was not. In order to compare effects of doing pre-processing and normalization at a later stage, a control analysis was first performed in which the data sets were run through the CSD framework without any pre-processing as a negative control for testing pre-processing effects. This data set consisted of 24,753 genes and will be referred to as **Analysis 0**. After loading in the thyroid cancer expression data into R, the filtering procedures were done in two different ways. As mentioned above, all genes who's values were less than 6 raw read counts in more than 20 % of samples were removed. After loading this data into edgeR, it was used to convert the raw reads into "count per million". Then one version of the thyroid cancer data was filtered identically as the healthy control data from GTEx and the other version of the thyroid cancer data was filtered according to edgeR's published recommended pipeline for differential analysis.

The first filtering strategy was a complete replication of all filtering step performed in the pre-processing method for the GTEX database. In this case, the expression vectors were transformed into TPM-units and filtered with a cutoff of 0.1 TPM in at least 20 % of samples, see step 3b in Fig. 3.2. This filtering process resulted in a data set of 16,728 genes (original data set size reduced by 32%). This version of the data set will be referred to as **Analysis 1**.

Because this filtering stringency resulted in the loss of 32% of the thyroid cancer data, the sick was pre-processed in a second way. Here, the normalization and filtering of content was done as recommended in the guidelines in a published paper presenting a workflow for differential gene co-expression with edgeR [105]. Expression data were required to have a "count per million"-value of at least 1 in at least two libraries, see step 2c in Fig. 3.2. Filtering CPM values rather than raw count values was to avoid favoring genes that are expressed in larger libraries. After the exclusion of genes with expression values below this criteria, the expression values were converted to TPM units, see step 3c in Fig. 3.2. This was a less stringent filtering threshold than implemented by GTEX and resulted in 20,657 genes in the data set (original data set size reduced by 17%). This version of the data set

will be referred to as **Analysis 2**.

Gene expression data from thyroid tissue of patients with thyroid carcinoma and healthy control subjects were used to generate a CSD network based on the method described in Chap. 2.9. The same workflow described in 3.1.4 was applied for Analysis 1 and Analysis 2. This resulted in four networks inferred on the basis if four importance values $10^{-4}, 10^{-4.5}, 10^{-5}$ and $10^{-5.5}$ for both Analysis 1 and Analysis 2.

### 3.2.2 Testing the effect of pre-processing strategies

In this thesis the role of proper data pre-processing was investigated by comparing differential networks from three different methods for pre-processing. The aim of this part was to examine the effect of alternative work-flows mainly differing in filtering cutoffs for low-count genes. All together, this resulted in three alternative ways of constructing differential co-expression analysis with normal expression data from GTEX and cancer expression data from the TCGA data-base . First, differential co-expression networks were generated for each alternative data set; Analysis 0 as negative control for filtering low count genes, Analysis 1 with identical pre-processing as the GTEx healthy data, and Analysis 2 with less stringent filtering as recommended by [105]. The four cut-off values used for network inference for each the three sets respectively were $10^{-4}$, $10^{-4.5}$, $10^{-5}$, and $10^{-5.5}$. In total 12 network were created. These were compared in several ways.

Networks were analyzed using the Cytoscape tool Network Analyzer [114]. Network metrics were gathered and presented in Table 5.1. The relative quantities of identical differential co-expression associations between the three analyses were computed. This measured the extent to which they resulted in similar differentially expressed genes, while taking into account the various interaction types genes could associate by. The effect of the three pre-processing methods on the contents related to biological function was investigated. The overlap coefficient was used to quantify the degree of overlap in the genes identified as network hubs between the analyses with correction for different sizes of the sets. This quantification was performed by selecting degree cut-offs for genes of $k \geq 10, 20, 30, 40$, and $k \geq 50$ separately for each analysis and compare each of them against each other. Results from this examination of hubs is presented in Figure 5.2. Lastly, the three analyses were assayed for ability to identify disease-associated genes as DEGs included in the networks. Results were compared between the three and formed a qualitative measure of inferred network success.

### 3.2.3 Parallel programming

When performing computations on large sets of data, like gene-expression data sets consisting of several thousand genes and hundreds of samples, the computational load becomes exceedingly large. Some processes implemented in the serial way can take longer to finish than is acceptable in the scope of a project time frame. By parallelization of computer programs,

computations may be run simultaneously. This reduces the total elapsed computation time needed, making it very well suited for e.g. correlations among genes in transcriptomic data.

To reduce the time span of a larger computational process of a script or program, implementing it in parallel is a way of separating the process into smaller entities, which then all can be run at the same time. Each of the processes a larger process is divided into are often distributed on the available computational processing units (CPUs) on a computer. Parallel computing relies on structuring a script or program to perform this computational delegation so that the parallelized processes are managed appropriately for effectiveness and success of the program's purpose. Distribution of input and output must be handled in the program in a way that facilitates rapid access and utilization of information across processors. Management of processes is important to ensure synchronization of processes for maximization of program speed. Implementation of parallelized programs can be done either by multi-processing, where each process is entirely separate from each other, of by multi-threading, where each thread may share information with others. Either type of parallel processing will only perform faster compared to programs implemented in serial if they are run on computer with hardware supporting parallelization [115].

In this thesis, parallel programming is both implemented in programs written in the programming languages *python* and in *R*.

### 3.2.4 Development of CSD framework with wTO

Weighted topological overlap (wTO), as discussed in section 2.6.3 was developed by Ravasz [78]. It is a similarity measure shown to capture biologically relevant associations between genes in differential co-expression networks. The current available Python-script developed by André Voigt had a run-time complexity of $O(n^3)$. For large data sets of 24,753 genes this would result in $O(n^3) = 15$ trillion calculations. With this run-time performance it's applicability to transcriptomic data where $n \approx 25,000$ is limited and proves impractical.

The original implementation of the script created an instance of a NetworkX [89] graph, which is a two-dimensional dictionary. The weighted sum of each gene's links to it's neighbors was computed by iteration through a double for-loop over this graph-object. For-loops may be too slow for a huge number of iterations, especially when the loops are double and the number of iterations required on each iteration layer is big. This implementation was thus improved upon by making the first layer of iteration parallelized. In the parallelized version of the script the outer layer of this double for-loop was separated into separate processes. Each of these processes had independent instances of the NetworkX-graph over which a single-layered for-loop computed the weighted sum of links for each node in the graph object. As each process needed a copy of the entire network to iterate through, the memory usage was extremely high. In stead of applying multi-processing, re-writing the script with multi-threading was done to improve the implementation. After running the multi-threaded Python-script for computing wTO for the differential co-expression network for almost 800

hours, it became clear that it would not finish within reasonable time. This could be due to computational overhang when multiple threads are trying to read information from one shared global variable.

As alternative strategy, the R-package *wTO* developed by Gysi et. al. [116] was employed. The approach to use the weighted topological overlap as basis for the CSD framework was first to do soft thresholding of the co-expression data, then applying the R-package *wTO* to transform the correlation to wTO, and lastly to bring this output back to a format fit for the CSD framework for network inference. This was done to both sets of expression data from thyroid cancer patients and healthy controls.

As an initial step before computing the wTO, soft thresholding was applied to the pairwise gene correlation values. This is applied to accentuate larger values of correlations among the many correlation values, while preserving the continuous nature of the co-expression measures. This also reduces the effect of noise in the gene expression values on the pairwise gene correlations, and is a biologically motivated criterion resulting in gene co-expression networks of scale-free topology [117].

Specifically, a continuous measure to assess their connection strength is used:

$$a_{ij} = s_{ij}^{\beta} \tag{3.1}$$

where $\beta$ is the thresholding parameter. High values $\beta$ will force low similarity towards zero whilst accentuating higher similarities. Here a value of $\beta = 5$ was selected to both maintain a scale-free network topology and retain the signs of correlation values.

The file containing the soft thresholded correlation values consisted of $n^2$ values for a data set of size $n$. This information was written independently for each gene pair on one line of the file. The input for the R-package *wTO* needed to be in matrix format, thus a Python script was written to do this conversion. So from the gene-pair correlation value file an analogous correlation matrix was written containing the identical information but in matrix format, together with a list of all the $n$ genes in the data set. This correlation matrix-file and the file with a list of gene names were read into R, in which the correlation matrix was stored as a data-frame for which the gene names were set as row- and column-names.

In the equation for weighted topological overlap (Eqn. 2.14) the connectivity of each node is computed, which is the sum of all weighted links connected to the node. In a signed network, there may be weight attributes of both negative and positive signs among the links a node has, which results in a cancelling affect in the node connectivity. To compute the real connectivity for the signed network, the absolute values of each weighted link is used to find this connectivity. This is ensured by setting the function parameter *sign = 'abs'* for the *wTO()*-function.

When this function was done running, the output matrix containing all wTO-values were multiplied with the signs of the correlation values. By letting these signs be the sign of the wTO, no node connectivity was potentially cancelled in the process of computation.

This method of preserving the sign of the co-expression measures (which is crucial to CSD method), addresses a major limitation in the original investigation of using wTO for similarity measure within the CSD framework [118], as observed in [119].

As a final step, a Python-script was written which converted this wTO-matrix back into a file with individual pair-wise genes and their wTO-value. Even though this workflow as a bit back-and-forth this enables the same variance estimation and network inference algorithms to be applied to the correlation matrix based on wTO. As already explained, there was motivation for expanding the alternative similarity measures available providing the basis of the CSD method for differential gene co-expression analysis - whilst keeping most of the robust network inference strategy integral.

This strategy worked well for large correlation-matrices, considerable computation time was still needed but at least computation of the wTO finished and the differential gene co-expression networks were generated. Results are presented in Chapter 5.2. This method was applied to both transcription data sets pairs of healthy controls and thyroid cancer patients, i.e. Analysis 1 (with stringent expression value-filtering, consisting of 16,728 genes) and Analysis 2 (with looser cut-off, consisting of 20,657 genes). For each of these two versions the identical sub-sampling algorithms was used to generate differential gene co-expression networks also for sub-samples of sizes 200, 100 and 50 random samples from the total set to investigate this similarity measure properties at low data set sample size. Results from the robustness analysis are found in Chapter 5.4.2.

### 3.2.5 Development of CSD framework with mutual information

One of the major aims of this thesis was to expand the range of different similarity measures forming the basis of CSD framework and potentially improve cluster analysis of differentially co-expressed genes. The linear similarity measure Spearman's rank correlation coefficient is employed in the standard way for calculation of pairwise gene associations. Thus, the introduction of a more general similarity measure capturing a wider range of association patterns was interesting to develop. Mutual information (MI) was chosen because it is a more general measure of similarity, capturing both linear and non-linear correlation patterns - a measure that potentially could capture new significant correlations between genes[61].

MI is based on the Shannon entropy, which was initially developed for discrete data, for measuring degree of independence between systems with a finite set of possible states [120]. To compute MI, the estimation of probabilities for each state is not straightforward when the data has an unlimited number of possible states. This is true for transcriptomic data, which are experimental measurements recorded on a continuous scale. For these the probability distributions are not known and need to be estimated. Thus, expression values must be discretized, usually by binning the values into discrete entities with histograms. Binning of continuous data is prone to systematic error [121], and thus various other estimation methods exist for computing mutual information.

The estimation method used in the R-package *parmigene* finds an estimate of MI from the average distance for $k$-nearest neighbors averaged over all values for a variable instead of binning the metrics. This algorithm find a distance to $k$ number of nearest neighbors in a two-dimensional joint space for a variables as a first step, and this calculated distance is used to find metrics at a distance strictly less than this average distance in the individual sub-space for the given variable. This is done to both variables in a pair for which MI is to be estimated. The MI for the two variables is generalized into higher dimensions by finding the maximum norm

$$||z - z'|| = max||x - x'||, ||y - y'|| \tag{3.2}$$

based on the space spanned by each of the two variables $X$, $Y$ and $Z = (X, Y)$. The distance between $z_i$ to its $k$ neighbor is calculated, which is the maximal distance between these points on the individual subspace of variable $X$ and $Y$. Neighboring point for the Z-subspace are ranked according to the distance $||z - z'||$ and metrics less than this distance are included. This algorithm is used to discretize the metric into units for which the joint probability distribution can be computed.

In on a comparison done on various estimators for MI as similarity measure for transcriptional regulatory network inference [122] found the network inference algorithm "CLR" included in the R-package *parmigene* [123] to be the best performing inference algorithm compared to many others. This study also found the estimation method of MI to be robust. The function in this package for estimation of MI was based on a $k$-nearest neighbor estimator. This MI estimator have minimal bias and is adaptive to yield high resolution estimates when there are numerous measurements [75].

The first implementation of the code composed of first calculating MI between all gene pairs in the transcriptomics data set as a first step and then converting this matrix into the pairwise text-format which is input for the variance estimation script from the CSD framework. Estimation of variance could then be done in the exact same way as for the the gene correlations computed with Spearman's correlation coefficient. By keeping as many scripts of the existing CSD workflow as possible was an important goal of this MI development part in order to keep sources or error minimal and maintain comparable results.

Instead of this computationally inefficient implementation, a vectorized version which could run in parallel was written using the *parmigene* R-package with $k$-nearest neighbor algorithms for MI estimation. First the expression-matrices for cancer and healthy patients was bootstrapped to estimate the variance in MI, an equal amount of samples as in the full set was drawn randomly for each bootstrap with replacement. This operation was ran in parallel 20 times to yield 20 matrices of MI-values between all gene pairs in the transcriptomic data set. The variance in MI for each pair was estimated on the basis of these 20 MI-values. The results from the parallel processing was stored in a list and split into separate lists for cancer and normal condition afterwards. These lists were combined into three-dimensional arrays for which the vectorized average could be computed. Each MI-value was multiplied with the sign of the correlation respective value. Then each of the C-, S- and D-scores were

calculated and the output was written in the same format as for the analogous from the correlation-based CSD framework. In this way the Python script which selects a significant subset of interactions based on a computed cut-off value for significance to be included in the inferred network was applicable on the output from the MI-code written in R.

Each of the 20 parallels in this R-program used the function *knnmi.all()* to estimate the MI for the transcriptomic data with *k=10* nearest neighbors. Each parallel was assigned to available cores on the computer. To address some initial memory usage challenges, the cluster types were set to type "FORK" in order for each of them to automatically contain all environment variables necessary for computation. This reduced memory usage and made the code feasible. For the smallest data set consisting of 16,728 genes, the entire computation time was 603 hours. This code was also applied to random sub-sets of the complete data set, which were generated with the same sampling method as for conventional CSD analysis. Generation of differential gene co-expression values for these smaller transcription data sets was performed identically as for the full set, by running multiple estimations of MI with the package *parmigene* in parallel. This generated differential co-expression values for all gene-pairs. The CSD framework for co-expression variance estimation and network inference based on four importance levels $p = 10^{-4}, 10^{-4.5}, 10^{-5}$ and $10^{-5.5}$ was applied as a last step. Results are presented in Chapter 5.3.

### 3.2.6 CSD framework with alternative Spearman's $\rho$

To investigate the role of variance correction in the nominator of the expressions for C-, S- and D-scores, an alternate version Python scripts calculating C-, S-, and D-scores was also created. In this version, differential co-expression networks are generated in a similar way except for the calculation of C-, S- and D-scores which are done without the variance in the nominator of the expressions. The variance is set to 1 for all these gene-pairs. Pair-wise genetic interaction differences and cut-offs for each score are computed in the same way as for the original implementation of the CSD framework.

This alternative version of Spearman's ranked correlation coefficient was motivated by the potentially valid exclusion of the variance estimation procedure described in 2.9.1. This is the most time-consuming part of the CSD framework. The same strategy of construction of differential co-expression networks for the full data set and smaller sub-sample expression sets was applied when using this alternate version of Spearman's rank correlation coefficient as well. This resulted in four networks, a full set and three sub-sample sets, for both the transcriptomic data of more and less stringent filtering, analyzed with the alternate correlation coefficient.

### 3.2.7 Comparison of alternative similarity measures

The four alternative similarity measures as basis for co-expression computations were compared in the aspect of their success as foundation for generation of biological meaningful

differential gene co-expression networks with high validity and robustness.

Networks for each of the four similarity measures were inferred with four importance levels each ($p = 10^{-4}, 10^{-4.5}, 10^{-5}$ and $10^{-5.5}$) with the CSD framework. These will be referred to by the following terms based on the similarity measure as basis for CSD:

- CSD: Spearman's ranked correlation coefficient $\rho$
- CSD-VAR: Spearman's ranked correlation coefficient $\rho$ without correction for variance in $\rho_{ij}$ for gene pair $i, j$
- wTO: weighted topological overlap
- MI: mutual information

The role of the similarity measures were examined by comparison of network characteristics. Network Analyzer [114] was used to extract relevant network parameters descriptive of network quality.

In order to compare the resulting networks generated on the basis of these different similarity measures in a quantitative way, their robustness for small sample-size was studied. The sample size will be denoted on the x-axis of the resulting comparison plots. For all categories of similarity measures applied, smaller random subsets from the expression data were made. Maximum number of samples corresponded to the entire data sets of both conditions, 504 with thyroid cancer and 399 normal controls. This was used as reference to compare the sub-sets with. Smaller sets were constructed by selecting a identical number of random samples from the thyroid cancer or normal data sets, and writing new test data-files with these sample sizes. For each data set a sub-sample of 200, 100 and 50 were made so that the smaller sets were contained in the bigger. The expression values for all genes for these sample-selections were included in the test-sets, constructing gene expression files of 200, 100 and 50 samples for both thyroid cancer and normal control transcriptomic data. The data set this analysis was performed on was Analysis 1, consisting of 16,728 genes. This effectively simulated low sample size. The same differential gene co-expression analysis workflow was applied to all these collections of sub-samples of transcriptomic data for thyroid cancer and the healthy control patients. The importance level for each of the inferred networks was $10^{-5}$. The ratio of the inferred interactions of the sub-sets that were also contained in the full set was quantified with the overlap coefficient. This was done analogously for networks based on the four alternative similarity measures to precisely quantify robustness to (simulated) low sample-size. Plots comparing the four similarity measures' robustness is presented in Chapter 5.4.2.

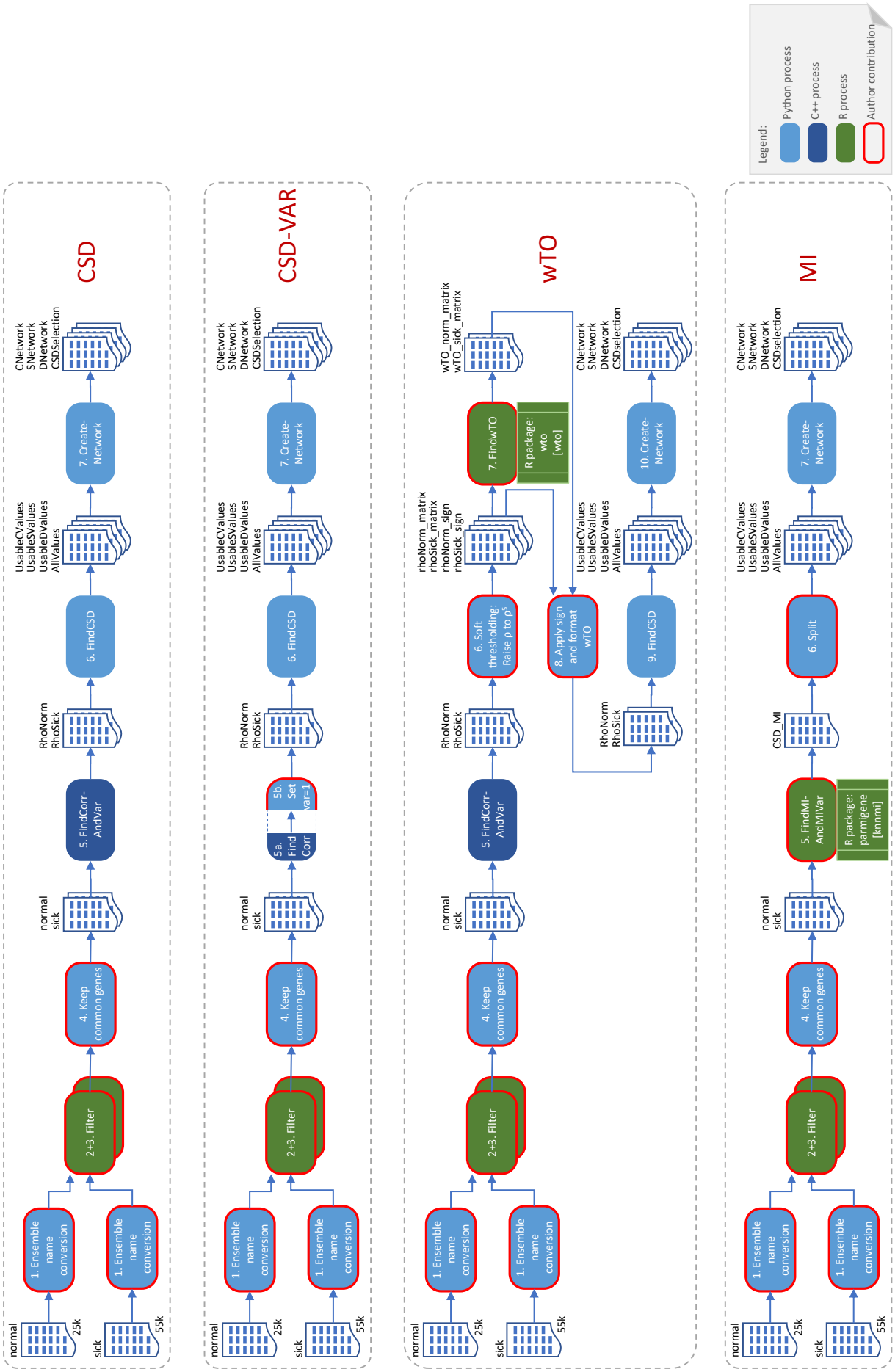


**Figure 3.3:** Illustration of the implementation of four alternative similarity measures to the CSD method.

Another way of comparison was to compare the network homogeneity. Box-plots of network homogeneity were generated for networks built upon each similarity measure and compared. This was done with the intent of examining how similarity measure related to the segregation of the various categories of changed transcriptional behaviour.

A last, but very important comparison of the similarity measures was to investigate their relative ability to identify thyroid cancer-associated genes. DEGs from networks based on each of the four similarity networks were converted to Ensemble IDs to maximize likelihood of the IDs to map to the GAD database [111]. This assay was performed with the PANTHER tool [83] and resulted in the number of thyroid cancer-associated genes present in each of the four networks, numbers of which were subsequently applied to assess the four similarity measures' ability to infer biologically informative differential gene co-expression networks.

# Chapter 4

# Results & Analysis: Application to Thyroid Cancer Expression Data

This chapter features an analysis of the differential co-expression networks constructed based on the transcripomic data of thyroid gland tissue from patients suffering from thyroid carcinoma and healthy controls. The differential co-expression networks were generated with the CSD framework. The goal was to use these networks for biological analysis of how thyroid cancer is manifested on gene co-expression patterns. These results may contribute to the elucidation of changed patterns in molecular mechanisms relevant for the pathogenicity this disease.

## 4.1 Construction of differential gene co-expression network

A differential gene co-expression network was constructed from the transcriptomic data set for thyroid gland tissue from thyroid cancer patients and healthy control persons. The CSD-method was employed to construct a network of genetic interactions. The data set consisted of gene expression measurements of 20,657 genes with 504 samples from patients with thyroid carcinoma and 399 healthy control persons. All gene expression samples were measured in thyroid gland tissue. The resulting network represents a significant selection of gene-pairs and their associations. With an importance level of $10^{-5}$ the CSD-method resulted in a network consisting of 1516 nodes (genes) and 3612 edges (gene-pair associations). The network has 303 C-links, 340 S-links and 873 D-links. As a general observation, this indicates that many of the inferred associations between gene expression patters are behaving in a differentiated manner between thyroid tissue with cancer and normal tissue. There are many genes which have a strong anti-correlated transcriptional pattern when the tissue changes from healthy to thyroid cancer.

The network is visualized in Cytoscape [107] in Fig. 4.1. Networks using four different importance levels were generated. With the intention of creating networks suitable for

analysis, an importance level resulting in a network of appropriate size was chosen. Those generated for $10^{-4}$ or $10^{-4.5}$ had an enormous amount of nodes (6593 and 3535 respectively) and that for $10^{-5.5}$ was perhaps small (694 nodes). Hence, the importance level $10^{-5}$ resulted in a network of desired size and density facilitating analysis.

The network consists of five bigger components with more than 40 nodes. The majority of nodes are found in three largest network components. The largest component is made up by 827 nodes (55%) and there are two smaller ones of 113 (7 %) and 101 (6%) genes respectively. The remaining 475 nodes are either in groups with a few nodes or just two in a pair.

Many of the links between genes in the largest connected component are differentiated and specific associations. The differentiated links are most numerous, and represent strong co-expression relations of genes to their neighbors which change sign when the condition changes. The specific links are strong strong co-expression relations which are completely condition dependent; for one condition they are strong and for the other there is no co-expression similarity to the neighbor genes no more. There are some occurrences of conserved associations as well, but the majority of links are differentiated and many are specific.

The topology of the largest component is interesting. It has two clear groupings with denser regions, forming almost two separate communities within this component. Between these two parts of the largest connected component there are some nodes linking them together. In the largest connected component the topological network structure is quite dichotomous, there are clear central regions of high interconnectivity but also presence of hubs linking to low-degree nodes and avoiding each other. In both of these dense communities in the largest component there are particularly interconnected neighborhoods and may indicate segregation of genes encoding molecules involved in different biological functions and potentially distinct cellular phenotypes [84].

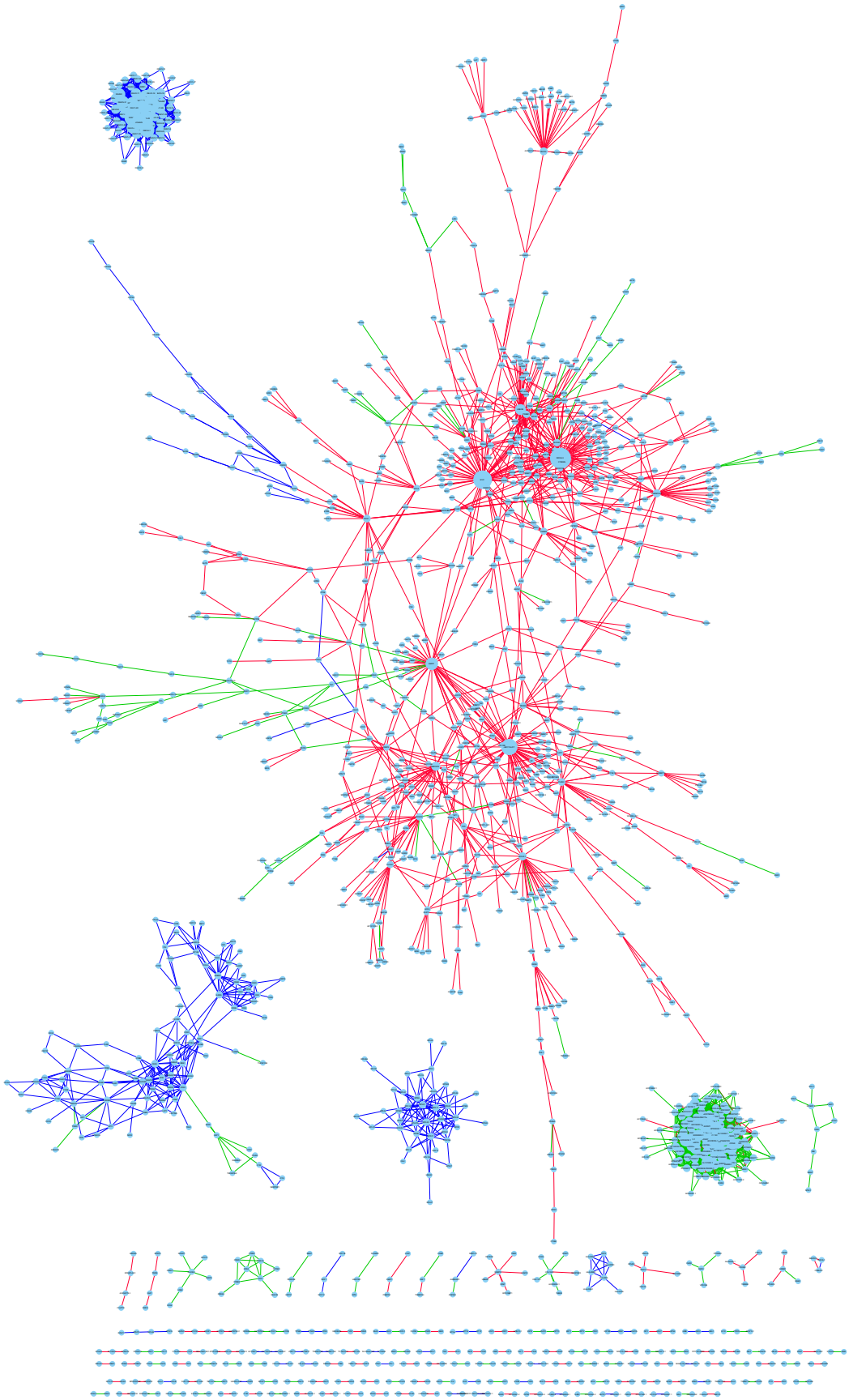

**Figure 4.1:** Plot showing CSD-network inferred with an importance level $10^{-5}$ on the data set with intermediate filtering process, Analysis 2, consisting of 20,657 genes. Inferred network consists of 1,516 nodes and 3,612 edges. Edges are coloured by type of interaction between the two compared conditions. Conserved links are blue, differentiated links are red and specific links are green. Node size is proportional to node degree.

## 4.2 Degree Distribution

The node degree distribution is a very important aspect of network analysis. It depicts the distribution of nodes with various degrees in the network and thus describes how complex the network structure is. In the differential gene co-expression network constructed the node degree distribution follows a power law, given by the equation $y = 524x^{-1.5}$. The correlation value of the fitted line is $R = 0.998$ ($R^2 = 0.887$). The degree distribution on a logarithmic scale is shown in Fig. 4.2.

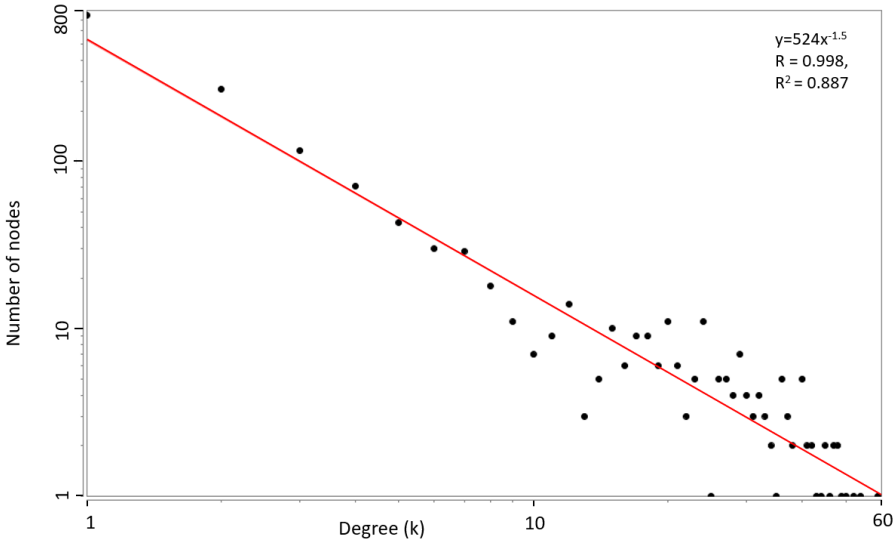


**Figure 4.2:** Plot showing degree distribution on a log-log scale for the CSD-network of Analysis 2, data set with 20,657 genes. Both axes are on logarithmic scale. The red line represents the function for the approximated power law fitted to the data points. It's expression is given in the top right corner.

This good fit between the power law-function and the data on a logarithmic scale indicates that the network is complex and scale free[1]. As is evident of the degree distribution there is a small number of nodes with very high degree while the vast majority of nodes have very few neighbors. The degree distribution indicates that there co-exists nodes of very low and very high degree in the network, being a characteristic of scale-free networks. This is very different from networks generated by random network models, which have a network topology characterized by an excess of nodes with comparable degree and no hubs. In scale-free networks such as this differential co-expression network, the hubs often have important metabolic roles. The low degree exponent of the network $\gamma = -1.5$ argues that the hubs are very central to the functionality of the thyroid tissue.

## 4.3 Hubs and assortativity

Hubs are defined as highly connected nodes, a loose definition which makes it expedient to set a limit to the degree of nodes which will be included in the definition as gene hubs in the inferred network. A limit of $k \geq 40$ produced a list of 22 genes with degree 40 or higher.

These correspond to the top 1.5 % of nodes in the network by degree. These high-degree nodes, the network *hubs*, can clearly be observed in several areas of the network in Fig. 4.1. These hubs are biologically interesting because they represent genes which co-express with a very high number of other genes - indicative of functional importance. Table 4.1 shows the network hubs sorted by degree and some of their associated properties, including information about the local clustering coefficient for each hub. Each row is coloured by the type of link most typical for the given hub, blue rows are hubs with predominantly C-links, green for S-links and red for D-links.

**Table 4.1:** Genes in the CSD-network with degree over 40, categorized as network hubs. $k$ denotes the node degree, $k_t$ is the number of link of type $t \in (C, S, D)$ for each hub. H = homogeneity. $C_v$ = clustering coefficient. Row colour describes the predominant link type for each hub respectively, blue if C-links, green if S-links and red if D-links.

| Name | $k$ | $k_C$ | $k_S$ | $k_D$ | H | $C_v$ |
|---|---|---|---|---|---|---|
| MICAL3 | 59 | 0 | 0 | 118 | 1.0 | 0 |
| EVC | 54 | 0 | 0 | 108 | 1.0 | 0 |
| IGLC3 | 52 | 104 | 0 | 0 | 1.0 | 0.45 |
| IGKV1-5 | 50 | 100 | 0 | 0 | 1.0 | 0.48 |
| FAM111A-DT | 49 | 0 | 0 | 98 | 1.0 | 0 |
| IGHV3-30 | 48 | 96 | 0 | 0 | 1.0 | 0.5 |
| IGHG1 | 48 | 96 | 0 | 0 | 1.0 | 0.5 |
| IGKV3-20 | 47 | 94 | 0 | 0 | 1.0 | 0.5 |
| JCHAIN | 47 | 94 | 0 | 0 | 1.0 | 0.5 |
| IGHG3 | 46 | 92 | 0 | 0 | 1.0 | 0.5 |
| IGKV4-1 | 45 | 90 | 0 | 0 | 1.0 | 0.5 |
| IGKV3-11 | 45 | 90 | 0 | 0 | 1.0 | 0.6 |
| IGLV1-40 | 44 | 88 | 0 | 0 | 1.0 | 0.5 |
| IGHA1 | 43 | 86 | 0 | 0 | 1.0 | 0.6 |
| AC097639.1 | 42 | 0 | 84 | 0 | 1.0 | 0.3 |
| AL121992.1 | 42 | 0 | 82 | 2 | 0.95 | 0.3 |
| AL009174.1 | 41 | 0 | 78 | 4 | 0.90 | 0.2 |
| AC034236.1 | 41 | 0 | 76 | 6 | 0.86 | 0.2 |
| AP004242.1 | 40 | 0 | 78 | 2 | 0.95 | 0.3 |
| MALAT1 | 40 | 0 | 72 | 8 | 0.82 | 0.3 |
| AC067735.1 | 40 | 0 | 80 | 0 | 1.0 | 0.3 |
| IGHV5-51 | 40 | 80 | 0 | 0 | 1.0 | 0.6 |
| AMOT | 40 | 0 | 6 | 74 | 0.86 | 0 |

We can observ that hubs with predominance of C-links have a substantially higher clustering coefficient than the average clustering coefficients for the network. Their co-expression associations remain intact between the studied conditions. C-dominated hubs with high clustering coefficients are likely to be found in well-defined cliques in the network where

mostly all neighbors are linked to each other [78]. Tightly interconnected clusters like these may form due to functional similarity. Their components are often vital for the integrity of the cellular behaviour, thus some are evolutionary conserved [124]. This testifies that the C-type network hubs may have essential roles in the cell which a are not changed in thyroid carcinoma. Conversely, the D-dominated hubs all have clustering coefficients of zero. These have numerous associations to other genes which switch sign between the studied conditions. These are likely to play a part in mediating the diseased phenotype in the thyroid tissue.

The gene with highest degree is the gene *MICAL3*, which is a protein-coding gene. It encodes a microtubule associated monooxigenase involved in microtubule filament disassembly functions in the cell by depolymerization. It is important for cellular processes, such as organization of the cytoskeleton and cell division. It is found in the largest network component in an especially densely interconnected region. The position of gene MICAL3 is indicated with an arrow in Fig. 4.3. It is a completely homogeneous gene, with associations to other genes of exclusively differentiated type. This hub has a very important function in the cell, and it's transcriptional behaviour is oppositely signed between thyroid cancer and normal tissue. This indicates that the co-expression pattern of this gene with all its neighbors changes significantly in thyroid cancer cells and the patterns of interactions of this microtubule monooxigenase with its neighbors' gene products is completely reversed.

*EVC* encodes the "Ellis-van Creveld syndrome protein", which is essential for signaling pathways involved in cell differentiation. It has been shown to stimulate thyroid cancer cell motility and invasiveness [125]. This gene is completely homogeneous; it always co-expresses with other genes oppositely between thyroid cancer and normal tissue. The last hub in the table, *AMOT*, is also a gene which protein product is involved in cytoskeletal organization [126]. The average betweenness centrality in the network is $\langle C_B \rangle = 0.074$. Both of these genes have a slightly higher betweenness centrality than most other hubs in the network, $C_B(EVC) = 0.4$ and $C_B(AMOT) = 0.3$. EVC and AMOT are both located in the outer regions of the two main community-like structures of the largest connected component in the network, indicated by arrows in Figure 4.3.

All the genes who's names begin with "IG" are immunoglobulin genes encoding antibodies, proteins that recognise antigens and important participants in the immune system. These hubs are marked with names and arrows to their location in the small network component in the top left of Fig. 4.3. There are exclusively conserved type associations in this component, indicating that co-expression patterns among these genes are similar in thyroid cancer and normal tissue. The structure is separated from the rest of the gene co-expression network, which also indicated that these genes form a functional module associated with normal immunity functions.

*FAM111A-DT* is a divergent transcript of the gene *FAM111A*. This transcript version encodes a long non-coding RNA-molecule. This gene is found in the largest network component, see Fig. 4.3. Mutation in this gene may cause disruption of thyroid gland function, like reduced parathyroid hormone production in the hypoparathyroid cells leading to im-

paired skeletal development [127]. The transcript has exclusively differentiated interactions to other genes in the network. The correlation in expression with all other genes it associates with completely reverses when going from normal thyroid tissue to cancerous. This may indicate that this long RNA-molecule is an antisense transcript of the *FAM111A*-genes promoter region, and may act as a gene expression regulator [128]. Many long non-coding RNA molecules have recently been shown to have important roles in cell differentiation [129] and are linked to development of cancer [130].

The hubs *AC097639.1, AL121992.1, AL009174.1* and *AP004242.1* are genes encoding mannosidases, a class of proteins involved with the degradation of the endoplasmatic reticulum (ER). These genes are found in the very interconnected component with many hubs on the lower right of Fig. 4.3. This component consists mainly of specific (green) links, but also have some differentiated (red) links. The gene *MALAT1* is also in this tight cluster. In thyroid tissue with carcinoma it has been shown that the expression level of *MALAT1* determines pro-apoptotic pathways in thyroid cancer cells and may trigger their death[131].

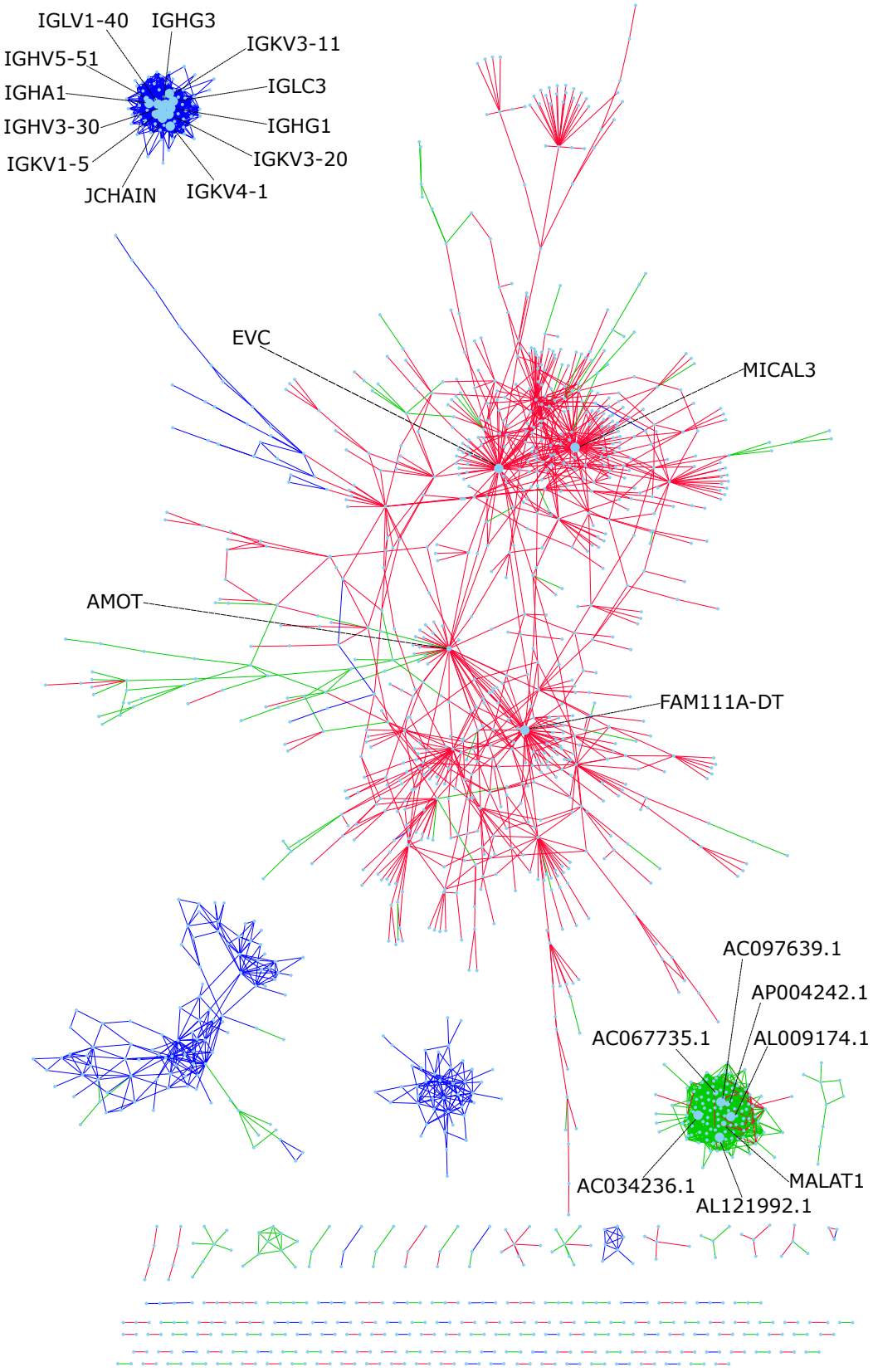


**Figure 4.3:** Visualization of the hubs of the CSD-network for Analysis 2, identical to that in Fig. 4.1. The links are coloured according to link type, C-links are blue, S-links are green, and D-links are red. Hubs (degree $k \geq 40$) are enlarged and their respective names and positions in the network are indicated by arrows.

## 4.4 Network homogeneity

Nodes in the network in Figure 4.1 associated with each other in many different ways. There were some variations in the distribution of association types for some genes, while for the vast majority the distribution of link types was narrower. These were dominated by specific interaction types and thus homogeneous in respect to associations. Node homogeneity-scores (H-score) calculated for all nodes sorted according to degree is presented in a box-plot in Fig. 4.4. This box plot is for the CSD-network based on Analysis 2, the intermediate stringency for expression vector filtration, consisting of 20,657 genes. The importance level of the network is $10^{-5}$. The homogeneity score (H-score) is quantified as a function of degree, and the distribution of H-scores for all nodes of a given degree is described by the properties of the box in the plot.

Network homogeneity analysis gave insight into the homogeneity of individual genes as well as the whole network. A node with one neighbor must have a H-score of 1 and a node with two neighbors can either have a H-score of 1 or 0.5, thus low-degree nodes in terms of homogeneity start to be interesting from $k = 3$. The distribution of link attributes between genes from Fig. 4.4 shows that for the majority of nodes their H-score means, the green squares, and their H-score median is found around $H \geq 0.9$. This trend is independent of degree. This shows that the genes in the network interact with their neighbors mostly of one single co-expression type.

Nodes of lower degree, $3 < k \leq 20$ have a predisposition for a specific type of interaction, but there are also some which do not. All of the squares denoting the means are found above $H = 0.9$, while the long whiskers of the box plots report that there are some cases of very low homogeneity. Because the degree of some of the nodes is very low here this is not so surprising, but the same long whiskers are indicated even for some nodes of high degree too. For nodes with $k = 25$ or $k = 28$ the homogeneity is a lot lower. Hence, there are some occurrences of genes with well-mixed interactions also.

There are several interesting nodes of intermediate degree and low degree of link-homogeneity. An example is CALR, which is a gene of lower degree that encodes calreticulin and has two C-links and three D-links. Calreticulin is often located in the endoplasmatic reticulum where it is involved in protein folding ensuring that new proteins are folded correctly. It has three D-links to the following genes: MAPKAPK3, which encodes a MAP kinase-activated protein kinase, which regulates many other functions of other proteins in response to stress, SPINT1, which encodes a protein that inhibits other proteins from degrading misfolded proteins, and DSC2, which encodes a transmembrane glycoprotein and who's reduced expression levels is associated with cancer (all this information is from the UniGene database). This demonstrates that genes with heterogeneous link distribution may be very informative in highlighting sites of transition form abnormal activity of important regulators of cellular activity which could provide new knowledge about the transcriptional changes characteristic of a studied phenotype.

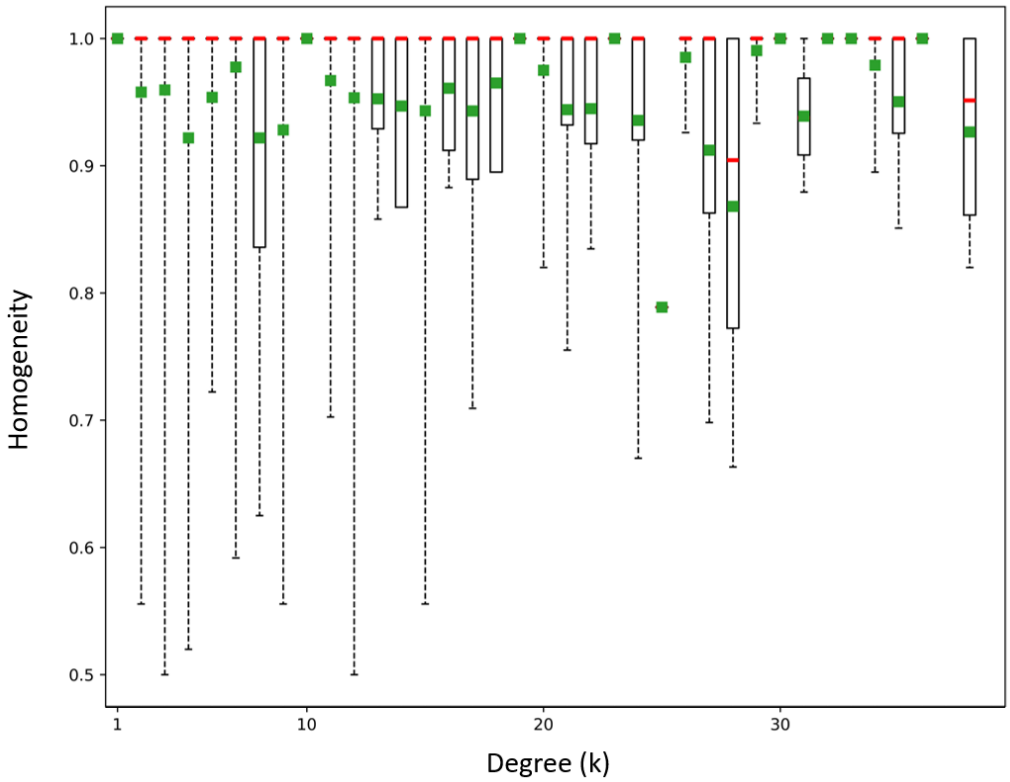


**Figure 4.4:** Box plot of network homogeneity for Analysis 2 network in Fig. 4.1 binned by node degree. Red bars correspond to the median of H, and the green squares denote the mean H. The top and bottom ends of the boxes represent first and third quartile (25th percentile and 75th percentile) respectively. The ends of the whiskers represent the minimum and maximum values of H for the given degree.

The network hubs with high degrees are generally more homogeneous. Most of the hubs, as also evident in Table 4.1 have close to or maximal scores for node homogeneity. As can be seen in the box plot there are also a substantial number of nodes of intermediate degree that tend to have relatively homogeneous neighborhoods with mostly interaction of one type.

The venn-diagram in Fig. 4.5 shows the summed number of interaction types and the combinations of their mixed types found for all nodes in the network. There are no nodes with all three different link-types, and very few nodes with mixed interactions types. There are substantially more interactions of differentiated type than conserved. If a gene does have more than one interaction type, it most often has both differentiated and specific links at the same time. Together, Fig. 4.4 and Fig. 4.5 imply that there is a high degree of node homogeneity in the network in general. This results in visually clear regions of nodes interconnected with principally one type of link, either C-, S-, or D-links. A high average homogeneity-level can be interpreted as a general tendency of genes to fall under one type of transcriptional regulation instead of many at the same time. Nodes being regulated by the same transcription regulator will be likely to associate with each other in typically one way, i.e. one type of (coloured) link. Transcriptional regulators may also be present in these homogeneous regions themselves. Some potential candidates of this are the network hubs, which are extremely homogeneous (see Table 4.1). Indeed, weakly bound transcriptional

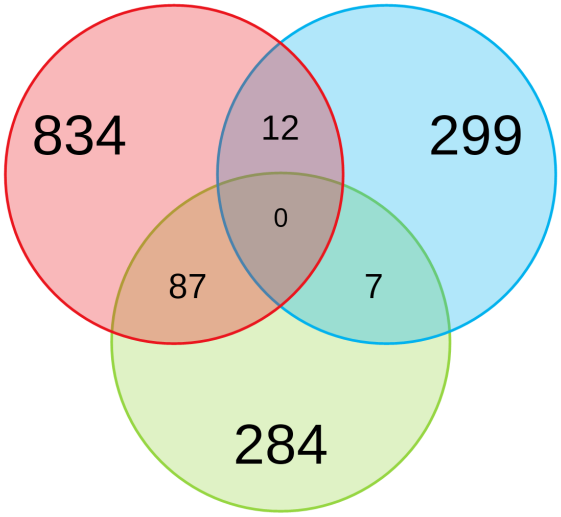


**Figure 4.5:** Venn diagram showing the mixing of differential co-expression types between the nodes in the network shown in Fig. 4.1. Blue circles contain the number of conserved links, green circles contain the number of specific links and the red circles contain the number of differentiated links, and their shared areas quantify the number of mixed interactions between the three types found in the network.

regulators contribute significantly to overall transcription in eukaryotes and are highly conserved [132]. Weak specificity of transcriptional regulators ensure robustness to mutations in their target genes. Weak-affinity transcription factor are essential for efficiency in mediating a change in transcription of a high number of genes rapidly to meet the constantly demands of cellular life [133]. Regions in the CSD-network with high average homogeneity and predominance of one link-type consist of nodes that correlate in transcription levels, and are thus likely to consist of genes being regulated by the same transcriptional regulators.

## 4.5 Biological process enrichment analysis

The differential co-expression network was investigated for enrichment of biological processes both for the complete graph and specific groups of nodes (genes). The results from this analyses shows the characteristic cellular processes the differentially expressed gene group are involved in. The result thus provided important information about which roles the genes have in the thyroid gland cells and if these functional roles segregated with the network neighborhoods. GO enrichment analysis for the whole CSD-network identified biological processes which are central in normal thyroid gland tissue but also may be indicative of malignant cellular behaviour.

GO functional enrichment analysis was performed on the differentially expressed genes in the CSD-network. This differential gene co-expression network with all link types C-, S- and D-links consisted of 1516 genes. Of these 1273 mapped to unique PANTHER gene IDs. Table 4.2 is a list of the biological processes enriched the most from the entire set of

differentially expressed gene in the network, sorted by the fold enrichment of the process in reference to a healthy human. The table shows that there is enrichment for many biological process that relate to the immune system. CD4-positive or CD8-positive alpha beta T cell lineage commitment is a process in which alpha beta developing T-cell commit into CD4+ or CD8+ lineage. Thymic antigen-presenting cells (APCs) use the major histocompatibility complex (MHC) to embed antigens which the alpha beta T-cells recognise with their antigen receptors. This processed is referred to as positive selection, while negative selection refers to recognition of self. This leads to skewed cellular fates for the T-cell and they may develop into regulatory T cells of differentiation [134]. The antigens connected to MHCs are thus important for orchestrating T-cell differentiation. MHC-genes are processed in the endoplasmatic reticulum (ER).

Table 4.2 also reports enrichment in process related to the ER and transport of proteins to the cell surface. SRP-dependent cotranslational protein targeting to the membrane is a process in which the signal recognition particle (SRP) is utilized to deliver proteins being transported from the ER after post-translational processing to the membrane [135]. Certain immune system pathways are normally involved in the process of producing and secreting thyroid hormone. Thyroid autoantigens like thyroglobulin, thyroid peroxidase, and the are TSH receptor are neccessary for the proper maturation of thyroid hormones [136]. Enrichment in associated pathways is thus not surprising . And exactly due to this dependency of the immune system and thyroid functionality makes it vulnerable to genetic mutations in or altered expression of the genes encoding proteins involved. The fact there is a clear dominance of differentiated links in this network (see Fig. 4.5) suggest that there is abnormal conduct of pathways of the immune system in thyroid cancer.

**Table 4.2:** Biological processes sorted by their fold enrichment identified by GO enrichment analysis based on all the differentially expressed genes of the CSD-network.

| GO biological process | FE | p-value |
|---|---|---|
| Positive thymic T cell selection | 10.5 | 3.6e-5 |
| Negative thymic T cell selection | 9 | 2.5e-4 |
| SRP-dependent cotranslational protein targeting to membrane | 8.34 | 4.6e-25 |
| T-helper cell lineage commitment | 8.25 | 1.1e-3 |
| Synapse pruning | 8.25 | 1.1e-3 |
| Renal filtration | 8.25 | 3.5e-4 |
| Positive regulation of RNA polymerase II transcriptional preinitiation complex assembly | 8.3 | 1.1e-3 |
| Regulation of humoral immune system process activation | 7.75 | 2.7e-26 |
| CD4-positive or CD8-positive, alpha-beta T cell lineage commitment | 7.7 | 1.6e-4 |
| Establishment of protein localization to endoplasmic reticulum | 7.53 | 6.6e-25 |
| Protein targeting to ER | 7.5 | 6.0e-24 |
| Nuclear-transcribed mRNA catabolic process, nonsense-mediated decay | 7.29 | 7.6e-25 |
| Protein folding in endoplasmic reticulum | 7.22 | 2.1e-4 |
| Fc receptor mediated stimulatory signaling pathway | 6.72 | 2.2e-24 |
| Fc epsilon receptor signaling pathway | 6.39 | 2.1e-28 |
| Humoral immune response mediated by immunoglobin | 6.36 | 2.9e-26 |
| Mitochondial ATP synthesis coupled proton transport | 6.29 | 1.6e-4 |
| Toll-like receptor 9 signaling pathway | 6.19 | 1.1e-3 |
| Antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent | 5.8 | 7.0e-11 |

In Table 4.3 the disease annotations identified for 873 of the 1308 (67%) genes mapping to the Genetic Association Database [5] are listed. The number of genes from the network that associated with a disease category are listed as "mapped genes." Fold enrichment is abbreviated by FE. The table provides and overview of the main disease-associations discovered in the full CSD-network by its composition of significant differentially expressed genes. All results presented are given with a corrected significance level for multiple comparison and only results with a false discovery rate less than 0.05 are included. All sicknesses are sorted by the number of mapped genes respectively.

The diseases associated with differentially expressed genes from the CSD-network charac-

**Table 4.3:** Diseases associated with the differentially expressed genes of the CSD-network identified with GO analysis, with the highest associated gene counts and their respective p-values and fold enrichment. Disease instances are sorted by the number of genes from the network mapping to the disease on GAD [5].

| Disease term | Mapped genes | p-value | FE |
|---|---|---|---|
| Acquired Immunodeficiency Syndrome | 88 | 1.0e-6 | 1.99 |
| Myocardial Infarction | 37 | 0.031 | 1.42 |
| Prostate cancer | 33 | 0.083 | 1.329 |
| Asthma | 31 | 0.061 | 1.38 |
| Longevity | 29 | 0.064 | 1.40 |
| Plasma HDL cholesterol (HDL-C) levels | 21 | 0.019 | 1.73 |
| Celiac disease | 19 | 1.0e-5 | 2.53 |
| Diabetes type 1 | 19 | 0.005 | 2.05 |
| Thyroid cancer | 18 | 1.0e-5 | 2.96 |
| Drug-related genes | 18 | 0.015 | 1.88 |
| Diabetes type 1 | 15 | 0.077 | 1.63 |
| Systemic lupus erythematosus | 14 | 0.042 | 1.83 |
| Cognitive trait | 13 | 0.012 | 2.26 |
| Aging/ Telomere Length | 13 | 0.013 | 2.24 |
| Rheumatoid Arthritis | 13 | 0.038 | 1.92 |
| Diabetes Mellitus Type 1 | 12 | 0.015 | 2.28 |
| Systemic Scleroderma | 10 | 1.0e-5 | 6.41 |
| Cardiomyopathy | 10 | 1.0e-5 | 4.89 |
| Lymphoma | 10 | 0.006 | 2.93 |

terize a system with strong enrichment of malignant cellular processes. There is indication of a combined role of both the immune system and cancer. The most relevant of these that categorize as diseases of the immune system are Aquired Imunodeficiency Syndrome (AIDS), asthma, celiac disease, diabetes, lupus, rheumatoid arthritis, and scleroderma. Enrichments for several types of cancer; cancers of the prostate, thyroid and lymphoma (blood lymphocyte cancer) are present. This disease-association enrichment seen in relation to Table 4.2 and Table 4.6 support the proposed interplay of the immune system and thyroid cancer responsible for its potentially invasive properties. Several genes of Table 4.6 are related to the enriched immune system pathways in Table 4.2 and are co-expressing in differentiated or specific manner relative to that of normal thyroid cells. The reversed and condition-specific transcriptional pattern in carcinogenic tissue compared to normal thyroid argues that the genetic interactions are systematically changed.

Table 4.4 shows the top ten genes in the network sorted by betweenness centrality. An interesting fact is that these genes also have relatively high degrees, and many of them are also listed as hubs in Tab. 4.1. Genes with high betweenness centrality have ability of monitoring communication between nodes in different regions of the network [55]. This

indicates that the hubs are very important for flow of information in the cell and likely act as key regulators of cellular activity.

In Table 4.5 the top ten genes sorted by eigenvector centrality are presented. Eigenvector centrality is used as an auxiliary quantification of how representative a gene is of the behaviour of it's immediate neighborhood collectively. Genes of high eigenvector centrality are influential in the network, meaning that they are connected to a high number of also well-connected genes. These gene then are likely to be very central players in the dynamic interplay of genes in the network and may be essential for regulation and synchronization of cellular pathways.

**Table 4.4:** Top genes sorted by betweenness centrality and their node degrees

| Gene name | Betweenness centrality | Degree |
|---|---|---|
| EVC | 0.106 | 54 |
| AMOT | 0.074 | 40 |
| FAM111A-DT | 0.065 | 49 |
| MICAL3 | 0.038 | 59 |
| BCL2L1 | 0.037 | 24 |
| WDR55 | 0.034 | 12 |
| DAG1 | 0.034 | 20 |
| PCYT1A | 0.030 | 5 |
| AC008870.2 | 0.028 | 4 |
| PLRG1 | 0.027 | 19 |

**Table 4.5:** Top genes sorted by eigenvector centrality and their node degrees

| Gene name | Eigenvector centrality | Degree |
|---|---|---|
| IGLC3 | 0.203 | 52 |
| IGHG1 | 0.202 | 48 |
| IGKV1-5 | 0.200 | 50 |
| IGHV3-30 | 0.198 | 48 |
| IGKV3-20 | 0.197 | 47 |
| IGHG3 | 0.193 | 46 |
| JCHAIN | 0.192 | 47 |
| IGKV3-11 | 0.192 | 45 |
| IGLV1-40 | 0.189 | 44 |
| IGHA1 | 0.184 | 43 |

## 4.6 Disease gene identification

Table 4.6 lists the 22 genes identified by DAVID as thyroid cancer-associated genes, their respective degree $k$, dominant link type $t \in (C, S, D)$, and the average shortest path length to other nodes in the network, denoted $\bar{d}_i$ for node $i$. These disease associated genes represent genes that have mapped to the Gene Association Database (GAD) [5] which related genes to specific diseases. The names of the differentially co-expressed genes were converted back to Ensemble IDs before submitting the GO query, while official gene symbols are used to name the genes in the table. Most of these genes had very low degrees and high average shortest path to other genes in the network, hence few of them were found in denser network neighborhoods but rather between them. The majority of the thyroid cancer genes interacted with other genes in exclusively differentiated manner, and many also had specific type links to neighbor genes.

**Table 4.6:** Genes in the CSD-network associated with thyroid cancer identified with GO analysis. The type denotes the predominant link type among a gene's associations ($t_{\in(C,S,D)}$). *IP3 = inositol 1,4,5-trisphosphate.

| Symbol | Gene name | $k$ | Type | $\bar{d}_i$ | Module |
|---|---|---|---|---|---|
| TPO | thyroid peroxidase | 12 | D | 5.50 | 5 |
| ITGB2 | Integrin $\beta$2 subunit | 6 | C | 3.15 | 7 |
| HLA-DRB1 | major histocomp. complex, class II, DR beta 1 | 5 | C | 1.67 | 45 |
| IRS1 | Insulin receptor substrate 1 | 4 | D | 5.60 | 5 |
| FN1 | Fibronectin 1 | 4 | S | 5.27 | 5 |
| CCL5 | C-C motif chemokine ligand 5 | 4 | C | 3.82 | 2 |
| STAT3 | signal transducer and activator of transcr. 3 | 4 | D | 4.52 | 5 |
| ITGA3 | Integrin subunit alpha 3 | 3 | S | 6.00 | 5 |
| TG | thyroglobulin | 3 | D | 5.98 | 5 |
| ITPR3 | IP3* receptor type 3 | 3 | D | 5.48 | 5 |
| ITPR1 | IP3* receptor type 1 | 2 | D | 7.84 | 5 |
| MATN2 | matrilin 2 | 2 | D | 6.64 | 5 |
| ADGRV1 | Adhesion G protein-coupled receptor V1 | 2 | D | 6.26 | 5 |
| TRIP12 | thyroid hormone receptor interactor 12 | 2 | D | 5.87 | 0 |
| HLA-DQA1 | major histocomp. complex, class II, DQ alpha 1 | 2 | C | 1.67 | 45 |
| S100A10 | Calcium binding protein | 2 | S | 1.33 | 5 |
| NCOA4 | nuclear receptor coactivator 4 | 1 | D | 8.23 | 5 |
| FAS | Fas cell surface death receptor | 1 | D | 7.58 | 5 |
| DIO1 | Deiodinase, iodothyronine type I | 1 | D | 7.36 | 5 |
| MAPK1 | mitogen-activated protein kinase 1 | 1 | D | 6.29 | 0 |
| SLC26A4 | Solute carrier family 26 member 4 | 1 | C | 1.00 | 1 |
| LGALS3 | Galectin 3 | 1 | S | 1.00 | - |

The genes identified in the differentially co-expression network as thyroid cancer-associated genes were mostly co-expressed in a differentiated or specific manner. This means that the co-expression of these genes in relation to genes they normally share close transcriptional behaviour with is completely reversed, which is manifested as red differentiated links in the network. The disease-associated genes were also often linked by specific type association to other genes, meaning that their transcriptional co-expression patterns are normally not found in thyroid tissue suddenly form in cancerous tissue or that normal co-expression patterns are lost. Most of these genes had very important roles in the regulation of cell cycle and thus their aberrant expression is likely to cause cancerous cell behaviour. Their biological roles associated mostly with the immune system or with cellular proliferation and life cycle regulation.

The THCA-associated genes are of paramount importance for both the differential gene co-expression analysis of thyroid carcinoma and for the comparison of different similarity measures' performance. They will thus be listed in A.1 together with information about their biological function and associations. An illustration of the disease genes in the CSD-network are given in Figure 4.6. Here the disease genes are indicated with enhanced node size.

The largest connected component of the CSD-network contained the highest number of THCA-associated genes. Here many of them were found in between denser network regions, and as seen in Table 4.6 some were characterized by very high average paths. This means that these are forming connections between different functional modules, implicating that these THCA-genes act as regulators of cellular function whose aberrant conduct communicate the wrong signals to the cell. Examples of such THCA-genes are TPO, ITPR3, FN1, STAT3, and IRS1.

*TPO* is found in the largest connected network component and interacts in opposite manner between the studied conditions. It is essential for thyroid hormone synthesis and a prognostic marker of thyroid cancer. It's differentiated expression levels has been related to the tumor development into invasive or indolent tumor sub-types [137].

*ITPR3* encodes a receptor for inositol 1,4,5-trisphosphate, which regulates the release of calcium from the endoplasmatic reticulum into the intracellular space.

*FN1* encodes fibronectin, and has four specific type links. Fibronectin 1 is involved with cell migration and metastasis. It is connected to one of largest network hubs AMOT. Through the condition-specific interaction with AMOT, abnormal correlations between genes that regulate cell migration and cytoskeletal organization appear, which could be responsible for thyroid cancer invasiveness. This supports the suggested prognostic determinant property of aggressive thyroid cancer [138].

*STAT3* belong to a family of genes, the STAT kinase family, which are transcriptional regulators. STAT3 correlated with other genes in a reversed manner between the studied condition. This indicates that it regulated transcription oppositely in thyroid cancer and

the normal case, which indicates this it could be responsible for inhibition of apoptosis and tumor cell proliferation [139].

*IRS1* is normally expressed at high levels in the thyroid. Here, it is D-linked to four neighbors, and through its neighbor USP3-AS1 these genes form a branch protruding out from the giant connected component where all co-expression associations are differentiated between normal and carcinogenic tissue.

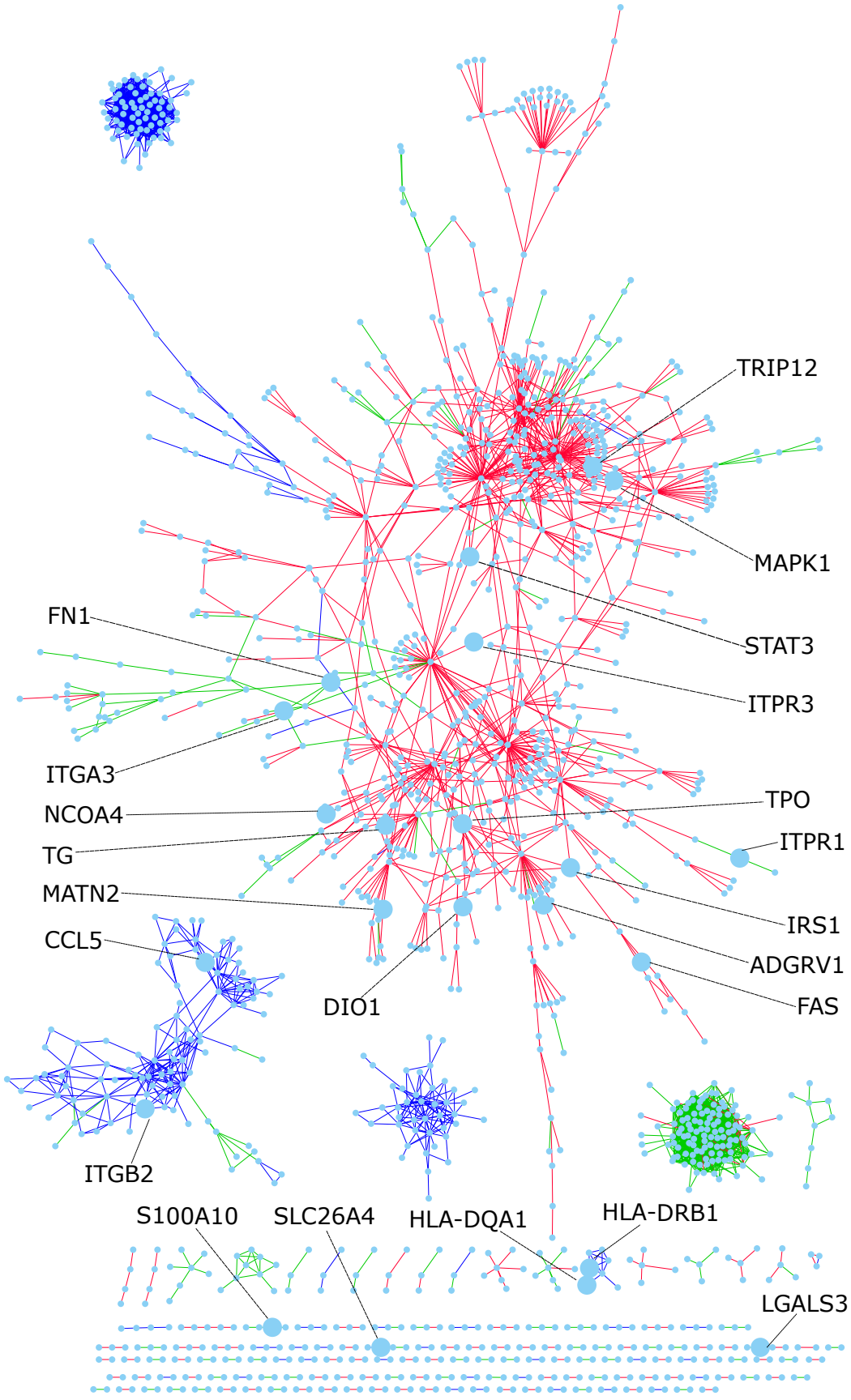


**Figure 4.6:** Illustration of CSD network for Analysis 2 with 22 thyroid-cancer genes identified as DEGs showed as enlarged nodes. This network is identical to that in Fig. 4.1. Thyroid-cancer genes are highlighted by name and arrows indicate their positions in the network. Edges between nodes are coloured by type of interaction between the conditions, conserved links are blue, differentiated links are red and specific links are green.

## 4.7 Investigation of network modules

Modules in the differential co-expression networks were investigated. Analyses with differential expression of human transcriptomic data often identify gene modules which are consistently expressed across many different samples of transcriptomic data. Such network neighborhoods are often dense and enriched for genes encoding proteins with general cell maintenance functions involved in regulation of the cell cycle, immune responses, extracellular matrix, and stress responses. But because misregulation of these processes are drivers of carcinogenic pathways, modules with enrichment for these processes in combination to high prevalence of differentiated and specific co-expression types will potentially distinguish dysfunctional cliques acting as disease modules.

Another strong motivation for module investigation was the attempt to find the thyroid cancer-associated genes from Chapter 4.6 listed in Table 4.6. The relative distribution of these disease genes was sought out in the modules in the pursuit of disease-associated cliques in the networks.

The Louvain algorithm implemented in Python [4] was employed for network community detection. The community detection was performed using its "best partition" functionality, as described in section 3.1.8. The outcome of this algorithm was 156 modules in the CSD network. Of these, there were 51 modules consisting of more than two genes investigated further. Table 4.7 shows these modules and Fig. 4.7 illustrated the partitioning of the modules in distinct areas of the differential co-expression network. In Table 4.7 the module number and the number of genes in each of them are listed. Also the average degree, $\langle k \rangle$, the average betweenness centrality $C_B$ and the clustering coefficient $C$ for the genes in the module is provided.

**Table 4.7:** Network modules with highest number of genes, their average degree, average betweenness centrality and average clustering coefficient.

| Module | #genes | $\langle k \rangle$ | $C_B$ | $C$ |
|---|---|---|---|---|
| 5 | 358 | 2.57 | 0.0021 | 0.004 |
| 0 | 310 | 3.10 | 0.0020 | 0.000 |
| 4 | 113 | 20.1 | 4.81e-5 | 0.285 |
| 11 | 69 | 21.9 | 2.25e-5 | 0.729 |
| 7 | 60 | 5.40 | 0.0001 | 0.416 |
| 34 | 55 | 2.67 | 0.0012 | 0.049 |
| 6 | 42 | 5.29 | 2.77e-5 | 0.390 |
| 14 | 42 | 2.14 | 0.0017 | 0.000 |
| 9 | 41 | 1.98 | 0.0016 | 0.000 |
| 2 | 32 | 4.78 | 9.11e-5 | 0.465 |
| 91 | 21 | 2.38 | 0.0015 | 0.130 |
| 21 | 9 | 3.56 | 2.42e-6 | 0.496 |
| 26 | 9 | 2.11 | 0.0001 | 0.259 |
| 78 | 9 | 2.00 | 5.23e-6 | 0.000 |
| 24 | 7 | 1.71 | 2.37e-6 | 0.000 |
| 31 | 7 | 1.71 | 1.87e-6 | 0.000 |
| 45 | 7 | 3.71 | 9.97e-7 | 0.738 |
| 96 | 7 | 1.71 | 1.87e-6 | 0.000 |
| 35 | 5 | 1.60 | 1.74e-6 | 0.000 |
| 42 | 5 | 1.60 | 1.05e-6 | 0.000 |

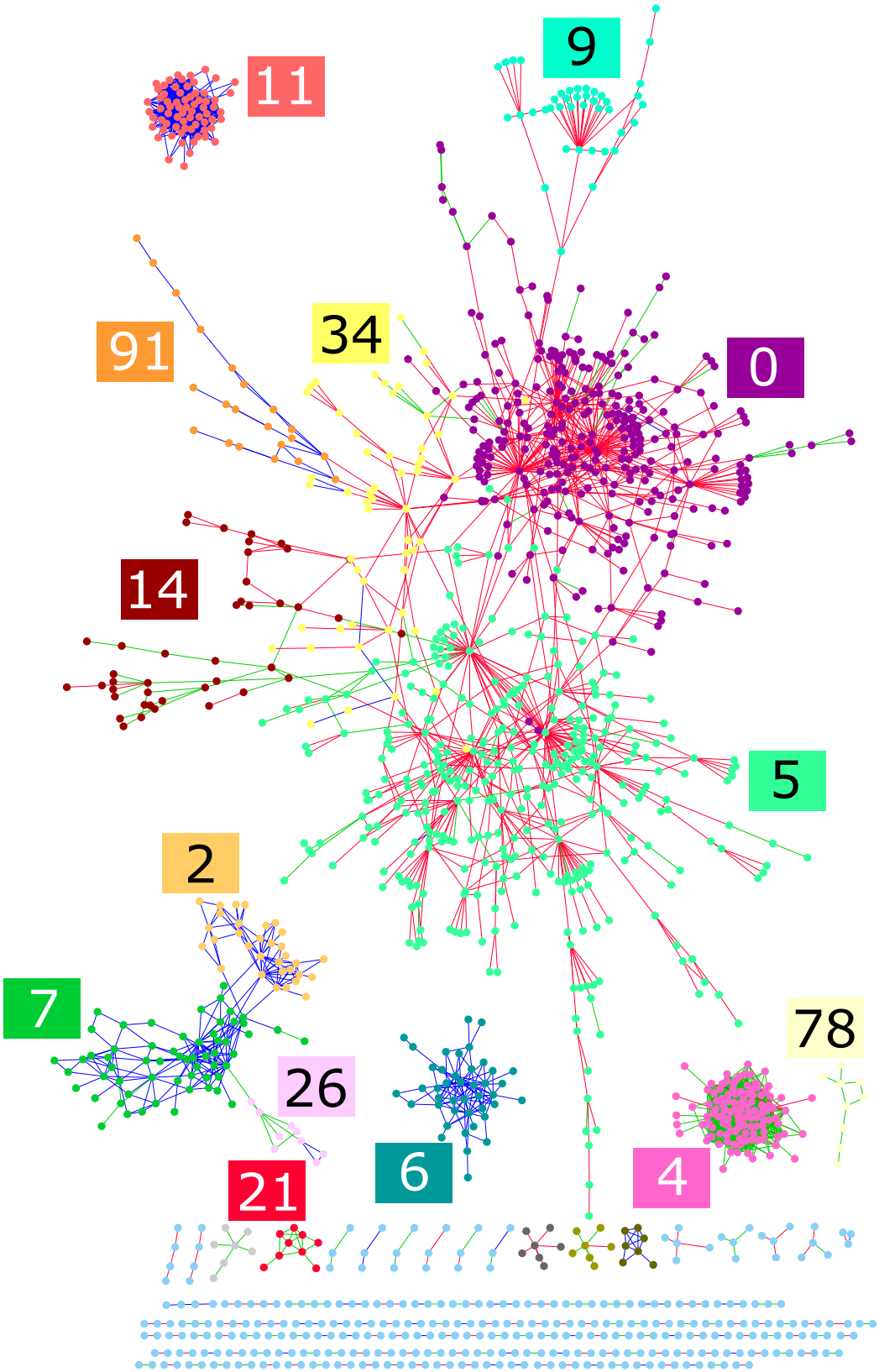


**Figure 4.7:** Plot showing communities in the differential gene co-expression network detected by the Louvain community detection algorithm [4]. Numbers on the figure denote the module number from Table 4.7 for each module respectively. Each community is coloured differently so that all gene nodes within the same community have the same colour. Edges between nodes are coloured by type of interaction between the conditions, conserved links are blue, differentiated links are red and specific links are green.

### 4.7.1 GO biological process enrichment analysis of modules

The following tables present the most highly enriched biological processes identified by using the GO tool PANTHER [110] with the differentially expressed genes from the most relevant network modules. All values in tables listing biological process enrichment identified by PANTHER have a statistically significant enrichment (false discovery rate (FDR) $P < 0.05$), and exclusively results with over a ten fold enrichment for these biological processes are listed. For some modules, GO enrichment analysis did not produce in any significant results associated with the genes constituting the module. These will not be further examined in the GO enrichment analysis of the network modules.

**Module number 5:** The largest module in the differential gene co-expression network with all link types consisted of 358 nodes. This module has a central placement in the network, as can be seen in Fig. 4.7 where the nodes of the module have bright green color. This module constitute approximately half of the largest connected component and forms a clear division in it. Within this module there are two types of links, specific and differentiated. The vast majority is specific. This indicates that co-expression relations between nodes in this module are likely to rooted in the studied condition; thyroid carcinoma.

GO enrichment analysis of the genes in this module identified 341 of the genes (95,3%). Of these 341 mapped IDs, 71 (20.8%) mapped to the OMIM database and 130 to KEGG pathways. The Table 4.8 shows enrichment for biological processes related to the genes of this module. The enriched processes show an interesting combination of thyroid hormone synthesis, carcinogenic processes, and immune responses possibly related to blood sugar (insulin/glucagon levels).

An interesting fact was that the functional enrichment mapped a substantial number of genes from this module to the OMIM database [140] and identified significant enrichment for thyroid cancer-related genes, with p-value = $4.04 \times 10^{-6}$, fold enrichment = 0.004 (bonferroni corrected p-value = 0.003 and FDR = 0.006). This module had the highest number of thyroid cancer-associated genes present, a total of 14 disease-genes out of the 22 (64 %) in Tab. 4.6 were found here. This is a strong indication that this module may actually represent a disease module.

**Module number 0:** The second largest module in the network consisted of 310 nodes, of which 293 mapped to the GO database. This module is coloured in purple in Fig. 4.7 and is found at the upper region of the large connected component. Table 4.9 lists pathway enrichment for genes of this module. Many of the enriched processes are involved in transcriptional or translational regulation. High fold enrichment of the processes are indicative of a clear functional clustering within this module. Two disease modules were found within this module, TPO and MAPK1.

**Module 7:** This module is found at the left of Fig. 4.7 in the second largest network component. The nodes belonging to this module are coloured green. Each of the identified GO biological processes have over a 50 fold enrichment. It should be notes that this module

**Table 4.8:** Table of the enriched KEGG pathways associated with Module 5, the largest module of the CSD network, identified by DAVID. The pathways are sorted by p-value. "benjamini" = Benjamini corrected p-value.

| KEGG pathway | p-value | benjamini |
|---|---|---|
| Metabolic pathways | 1.3E-3 | 2.5E-1 |
| Thyroid hormone synthesis | 1.9E-3 | 1.9E-1 |
| Small cell lung cancer | 5.1E-3 | 3.0E-1 |
| Aldosterone-regulated sodium reabsorption | 5.8E-3 | 2.7E-1 |
| Glucagon signaling pathway | 1.0E-2 | 3.6E-1 |
| Pathways in cancer | 1.5E-2 | 4.2E-1 |
| Insulin resistance | 1.6E-2 | 3.8E-1 |
| Proteoglycans in cancer | 3.4E-2 | 6.0E-1 |

**Table 4.9:** Over-represented biological processes associated with module number 0 of the CSD network as identified by GO, sorted by fold enrichment.

| GO biological process | Fold enrichment |
|---|---|
| Autophagy of peroxisome | 43.00 |
| G-quadruplex DNA unwinding | 30.71 |
| COPII-coated vesicle cargo loading | 22.05 |
| Pos. regulation of cytoplasmic translation | 21.50 |
| Regulation of histone ubiquitination | 21.50 |
| Proteasomal ubiquitin-independent protein catabolic process | 14.33 |
| Regulation of mRNA 3'-end processing | 13.27 |
| Tricarboxylic acid cycle | 11.94 |

had an interesting profile of biological process enrichment even though all gene-pair co-expression associations are of conserved type. The enriched biological process of this node are related to the immune response. The genes of this module show a high level of functional clustering. One thyroid cancer-related gene, ITGB2, belonged to this module.

**Module 34:** This module is found at the left in Figure 4.7, coloured in yellow. It consists of 55 genes which are loosely bound in the largest connected component of the network. It's average degree is low but the betweenness centrality $C_B = 0.0012$ is high compared to some other modules of similar size. of the largest connected component of the network in a dark green-blue colour. It is predominantly composed of nodes interacting in differentiated manner (red D-links) between the two studied conditions. GO enrichment analysis results are listed in Table 4.11, where 16 GO biological processes with over 20 fold enrichment are included. These processes show that genes in this module are involved with many important regulatory processes, and may act as bridges connecting different cellular pathways. The variety in pathways enriched in this module advocate such a role.

**Table 4.10:** Over-represented biological processes associated with module number 7 of the CSD network as identified by GO sorted by fold enrichment.

| GO biological process | Fold enrichment |
|---|---|
| Pos. reg. of hippocampal neuron apoptotic process | > 100 |
| Pos. reg. of type I hypersensitivity | > 100 |
| Pos. reg. of type III hypersensitivity | > 100 |
| Pos. reg. of prostaglandin-E synthase activity | > 100 |
| Vertebrate eye-specific patterning | > 100 |
| Complement-mediated synapse pruning | > 100 |
| Pos. reg. of microglial cell mediated cytotoxicity | > 100 |
| Pos. reg. of neutrophil degranulation | > 100 |
| Neg. reg. of dopamine metabolic process | > 100 |
| Pos. reg. of microglial cell activation | 88.97 |
| Microglial cell activation | 80.88 |
| Respiratory burst | 67.78 |
| Pos. reg. of interleukin-4 production | 61.89 |
| Pos. reg. of B cell differentiation | 59.31 |
| Macrophage activation involved in immune response | 59.31 |

**Table 4.11:** Table of the 16 top over-represented biological processes associated with module number 34 of the CSD network as identified by GO sorted by fold enrichment.

| GO biological process | Fold enrichment |
|---|---|
| Sequestering of calcium ion | > 100 |
| Pos. reg. of oxidative stress-induced intrinsic apoptotic signaling pathway | > 100 |
| Bundle of His cell-Purkinje myocyte adhesion involved in cell communication | > 100 |
| Protein folding in endoplasmic reticulum | > 100 |
| Mitochondrial ATP synthesis coupled proton transport | 54.54 |
| Cristae formation | 36.94 |
| Reg. of cellular amino acid metabolic process | 36.36 |
| Antigen processing, presentation of exogenous peptide antigen via MHC class I, TAP-dependent | 35.16 |
| Mitochondrial electron transport, NADH to ubiquinone | 31.16 |
| Reg. of hematopoietic stem cell differentiation | 28.88 |
| Substantia nigra development | 24.90 |
| Reg. of transcription from RNA pol. II promoter in response to hypoxia | 41.66 |
| NIK/NF-kappaB signaling | 23.56 |
| Mitochondrial respiratory chain complex I assembly | 23.14 |
| Negative regulation of G2/M transition of mitotic cell cycle | 21.45 |
| SCF-dependent proteasomal ubiquitin-dependent protein catabolic process | 20.98 |

# Chapter 5

# Results & Analysis: Method Development

The main aim of this chapter is present the outcomes of the method development part of this master thesis. First, this chapter will compare outcomes resulting from different pre-processing alternatives in the aspect of their quality and influence on their outcomes on differential co-expression analysis. Second, comparisons of differential gene co-expression based on four different similarity measures will be presented. These results were generated from the four different work-flows presented in Figure 3.2.7. The aim is to elucidate differences rooted in similarity measures between the inferred networks. In the investigation of both the roles of pre-processing and similarity measures, assessment of quantitative and qualitative properties will be performed.

## 5.1 Effect of pre-processing methods

The effect of pre-processing was investigated by producing three different combinations of normalized thyroid cancer data sets. Refer to Figure 3.2 for illustration. Networks of differentially co-expressed genes were constructed for each of these three data sets with alternative pre-processing:

- ANALYSIS 0 (An.0) : Converting ensemble gene IDs to official gene symbols only. Expression data sets were filtered to contain a common set as in both normal and thyroid cancer transcriptomic sets. Size: 24,753 gene expression vectors.
- ANALYSIS 1 (An.1) : Same as for An.0., but included removal of any gene who's raw count vector for all samples has less than 6 reads for more than 20% of samples. Reads were upper 75%th-quantile library scaled and genes who's TPM-values are less than 0.1 for more 20 % of samples are excluded. Units converted to TMM values with the R-package edgeR. Lastly, an inverse normal transform was applied. Size: 16,728 gene expression vectors.

- ANALYSIS 2 (An.2) : Similar first step as for An.0. Count values were converted to CPM values with edgeR, and exclusively genes who's CPM values were more than 1 in both libraries were retained. Reads were upper 75%th-quantile library scaled and an inverse normal transform was applied. Size: 20,657 gene expression vectors.

The effect of proper pre-processing and quality control was explored by comparing the networks resulting from differential gene co-expression (DGCE) analysis using the CSD framework with these three sets of transcriptomic data. In each of them the normal transcriptomic data is identical because it was already normalized when accessed and downloaded. The three different strategies for pre-processing - consisting of both gene filtration and expression level normalization - were performed on the data downloaded from TCGA to look into the possible outcomes on a differential co-expression analysis.

For each variation of pre-processing method the corresponding data set was used to infer networks with the CSD-framework on the basis of four importance values; $p = 10^{-4}, 10^{-4.5}, 10^{-5}$, and $p = 10^{-5.5}$. The network parameters of the resulting four differential co-expression networks for each analysis is listed in Table 5.1. The size of the data set in each analysis is listed under each of the three sets. The importance level $p$ is the common importance level used to estimate threshold values $X_p^{C,S,D}$ for each of the association type scores in order to map them in a comparable scale [3]. Genes and edges correspond to the number of genes and edges in the networks respectively, $\langle k \rangle$ is the average degree, $C$ is the average clustering coefficient. $\bar{d}$ is the average shortest path length in the network and the power law parameters $a$ and $b$ are constants from a fitted power law function to the log-log plot of the networks' degree distributions given by $y = ax^b$.

**Table 5.1:** Network parameters for the networks inferred on the basis of four different importance values, each generated for data sets subject to different pre-processing strategies.

| Analysis: size | p | Genes | Edges | $\langle k \rangle$ | C | Diam. | $\bar{d}$ | Power law |
|---|---|---|---|---|---|---|---|---|
| An.0 : 24,753 | $10^{-4}$ | 6826 | 50303 | 14.7 | 0.12 | 15 | 4.2 | a = 7089, b = -1.7 |
| | $10^{-4.5}$ | 4289 | 16960 | 7.9 | 0.10 | 19 | 4.8 | a = 3975, b = -1.8 |
| | $10^{-5}$ | 2285 | 4954 | 4.4 | 0.08 | 16 | 5.0 | a = 1375, b = -1.8 |
| | $10^{-5.5}$ | 1179 | 1665 | 2.8 | 0.06 | 16 | 5.6 | a = 639, b = -1.9 |
| An.1 : 16,728 | $10^{-4}$ | 4700 | 24301 | 10.3 | 0.12 | 13 | 4.8 | a = 3146, b = -1.6 |
| | $10^{-4.5}$ | 2695 | 8292 | 6.2 | 0.10 | 22 | 6.4 | a = 1525, b = -1.7 |
| | $10^{-5}$ | 1313 | 2442 | 3.7 | 0.07 | 23 | 3.7 | a = 646, b = -1.7 |
| | $10^{-5.5}$ | 587 | 804 | 2.7 | 0.09 | 14 | 5.2 | a = 277, b = -1.7 |
| An.2 : 20,657 | $10^{-4}$ | 6593 | 32569 | 9.9 | 0.10 | 18 | 4.9 | a = 4170, b = -1.6 |
| | $10^{-4.5}$ | 3535 | 11709 | 6.6 | 0.10 | 16 | 6.1 | a = 1343, b = -1.5 |
| | $10^{-5}$ | 1516 | 3612 | 4.8 | 0.11 | 18 | 6.4 | a = 524, b = -1.5 |
| | $10^{-5.5}$ | 694 | 1213 | 3.5 | 0.1 | 23 | 7.0 | a = 330, b = -1.6 |

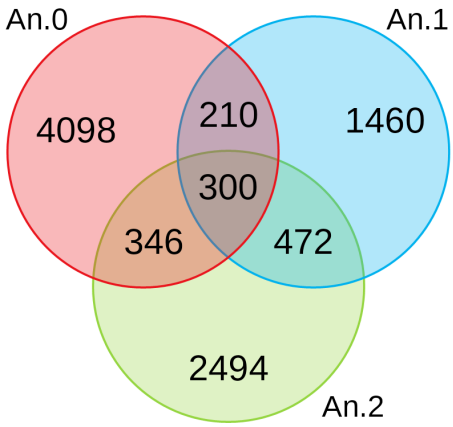


**Figure 5.1:** Venn-diagram of the distribution of common and unique inferred genetic associations between Analysis 0, Analysis 1, and Analysis 2. Comparison of networks inferred with importance value $p = 10^{-5}$ created with the CSD-method.

From these network parameters it becomes apparent that the data sets were somewhat different depending on the pre-processing applied. In terms of the number of genes included in the network as significant differentially expressed genes (DEGs) there were clear distinctions between the studied conditions. Except for the importance value $10^{-4}$, An.1 and An.2 appear more similar to each other than either of them seem to be compared with An.0. For $p = 10^{-4}$ in An.2 the network consists of approximately the same number of genes compared to An.0, but An.2 has a lot fewer edges. The networks of An.0 are systematically also more "efficient" in terms of a lower average shortest path. For most importance values it has lower $\bar{d}$ than the networks of the other analyses. This may however be attributed to the higher relative number of edges found here than in e.g. An.2 with the same number of genes.

Figure 5.1 depicts the relative amount of gene co-expression patterns shared between the different analyses in a venn-diagram. It shows how many of the links are exclusively found in each of them separately and how many are in common sets. Fig. 5.1 illustrates a higher similarity in identified differential gene association composition between An.1 and An.2 than each of these two has in relation to An.0 separately. An.1 and An.2 share 36% more gene associations than An.0 and An.2, reporting a higher compliance among genes identified as DEGs between the two studied conditions for data sets with more similar pre-processing.

The exceptionally high number of gene associations in An.0 may be an artefact of the co-expression network inference method based on raw RNA-seq data. In the thyroid cancer data set in An.0 there were several occurrences of genes who's expression vectors contained high percentage of zero-instances. This happens in RNA-seq transcription quantification when some genes' expression are unique to one condition and highly expressed for this condition. The sequencing sensitivity for other gene probes of the sample will thus be decreased and the data may become skewed [50]. Hence, this is not a weakness of differential gene co-expression as strategy to compare transcriptional patterns between carcinogenic and healthy tissue but rather a limitation of the RNA-sequencing technology. Consequently, many of these low-count genes become highly correlated with each other and have potential to be inferred as

DEGs when compared to transcription profiles from a set of samples representing a different condition. RNA-seq is thus prone to yield false positive rates in co-expression analysis when not accounting for expression profiles rich in zero-instances.

The potential effects of the pre-processing method of the RNA-seq data on the network structure and its components was also interesting. These were investigated by testing the similarities of the network hubs between the three analyses. For each of them their genes were sorted by degree and sets of genes over a certain degree cut-off were compared between them. Then the overlap coefficient of the content in each compared set was calculated in order to find a measure of similarity corrected for different sizes of the sets. Overlap study of hub sets between the three different analyses, for five different cut-off values for node degree was performed and results are presented in Figure 5.2. For each value along the x-axis the overlap coefficient is shown between two sets with a cross, for all three combinations of compared analyses. The overlap coefficient quantifies how many genes are found in both networks for a certain degree-threshold. This essentially compares the composition of hubs in the three networks and measure how many of the hubs are common.

The highest overlap coefficients of hubs between analyses indicated in Fig. 5.2 are found among An.1 and An.2. These are both analyses with pre-processing applied and seem to have several characteristics that are more similar than with the non-processed set An.0. The lowest scores of overlap are consistently observed between An.0 and An.2. This is not very surprising given the difference between them; An.0 with no filtering for low-count genes and An.0 with the strictest filtering of these.

A final assessment of effect of pre-processing procedure consisted of testing differences of the analyses in identifying thyroid cancer-associated genes. Here, associations between differentially expressed genes from all three alternatives An.0, An.1 and An.2 were used to query the functional enrichment tool PANTHER [110]. All queries were done based on networks of importance value $p = 10^{-5}$. This resulted in positive results for An.0. The original list of thyroid cancer-associated DEGs from Table 4.6 are based on An.2. The query did not yield any results for An.1, most likely because of lower size. But for An.0, which was the largest data set, 15 new thyroid cancer-associated genes were identified. There are listed in Table B.4. Most of the disease-genes were linked to their neighbors by specific or differentiated links. Some of the disease genes that were not in a pair, hence with $k \neq 1$ and $\bar{d} \neq 1$, had high values for average shortest path $\bar{d}$. These could have important functions in the cell mediating information from various regions of the network, potentially as regulators of transcription. The fact that these new thyroid cancer genes were obtained by the data set with no quality control of data does not necessarily show that proper quality control is not needed, but it does support performing differential gene co-expression with the CSD-framework in parallel for multiple variations of the same data set if possible as they may be able to discover different associations and disease-related genes.

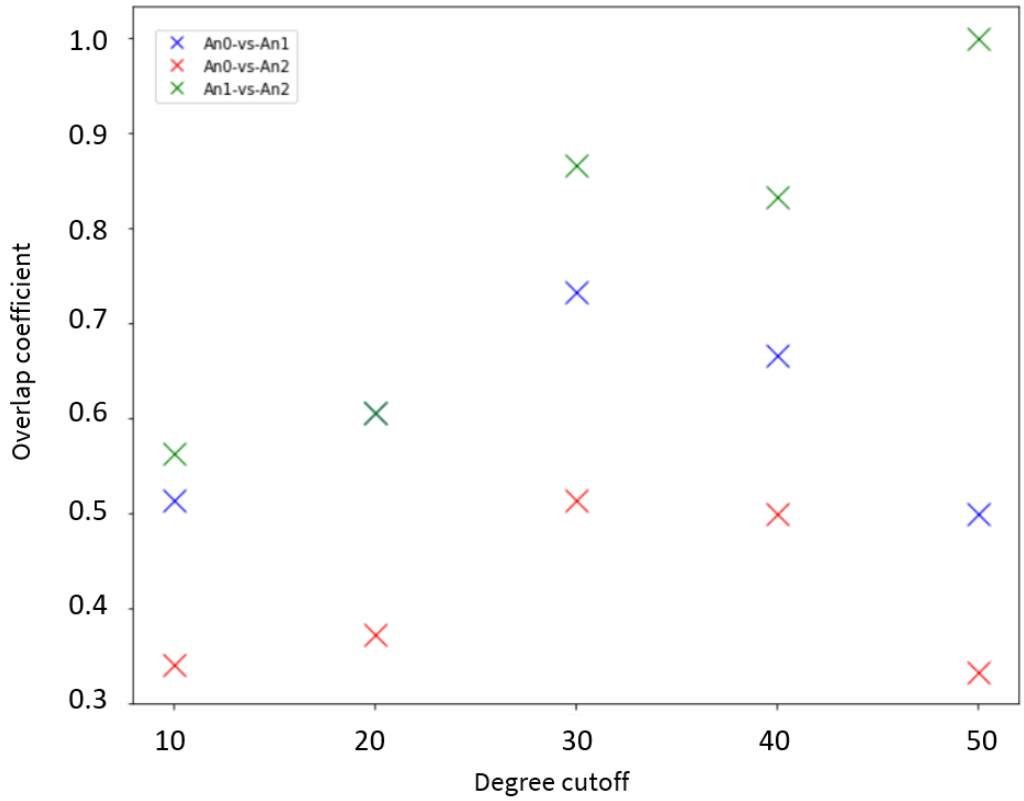


**Figure 5.2:** Plot of relative compliance between the three alternatively pre-processed data sets An.0, An.1, and An.2, in respect to identical content of nodes identified as hubs. Ratio of compliance among hubs is quantified with the overlap coefficient in sets of genes with increasing degree cutoffs, which increases along the x-axis. Increasing compliance increases along the y-axis. All three versions of the data set are compared to each other as indicated in the legend, and respective degree of overlap in each comparison is indicated with a cross in the plot.

## 5.2 Weighted topological overlap as similarity measure

Because the implementation of the python-script for calculation of weighted topological overlap measure had potential to be faster, an attempt to improve it was done. To reduce computation time the improvement of the program consisted of developing a parallelized implementation of the script. Both implementation with multi-processing and multi-threading were attempted but did not result in a faster implementation.

Instead, the R-package *wTO* developed by Gysi et. al. [116] was used to transform the co-expression data to weighted topological overlap-measures with soft thresholding with thresholding parameter $\beta = 5$ and the signed version of wTO. This procedure was performed to co-expression matrices of both normal and thyroid cancer transcriptomic data, and the output was successfully used to infer a differential co-expression network with the CSD-method.

Figure 5.3 shows the resulting CSD-network based on wTO inferred with importance level $10^{-5}$. The networks generated on the basis of weighted topological overlap as similarity measure generated networks with high ability of discrimination of genes involved in co-

expression patterns of the three association-types; conserved, differentiated, and specific. Visualization of the network with Cytoscape [107] clearly grouped network components according to distinct differential co-expression category. This is supportive of literature claiming that the topological overlap measure produces meaningful biological clusters of nodes in this type of networks [65].

The structure of this network was interesting. In Fig. 5.3 it can clearly be seen that the network is almost completely partitioned according to association type, with segregation of conserved, differentiated, and specific links in definite clusters. Table 5.2 lists network properties of this network, the number of genes and edges, average degree $\langle k \rangle$, clustering coefficient (C), diameter and average shortest path $\bar{d}$. It also provides corresponding information about a network inferred by the same importance value $p = 10^{-5}$ based on Spearman's rank correlation coefficient. The wTO-network also has a relatively smaller diameter and average shortest path compared to the $\rho$-based network. From comparison of data in Table 5.2 it is evident that the wTO-measure produces networks of higher interconnectivity - i.e. with higher average degree and higher average clustering coefficient.

**Table 5.2:** Table of network parameters for the networks based on WTO and and on Spearman's correlation coefficient, both with $p = 10^{-5}$.

| $S_{ij}$ | Genes | Edges | $\langle k \rangle$ | C | Diam. | $\bar{d}$ | Power law |
|---|---|---|---|---|---|---|---|
| wTO | 770 | 4831 | 12.6 | 0.27 | 11 | 3.5 | a = 190, b = -1.2 |
| $\rho$ | 1516 | 3612 | 4.8 | 0.11 | 18 | 6.4 | a = 524, b = -1.5 |

Another interesting observation is that this network had disassortative tendencies, especially in the second largest component consisting of predominantly specific type links and dense regions of tightly interconnected nodes with mostly conserved type links. This component is on the bottom left of Figure 5.3. The largest component consisting of virtually only differentiated type links, on the other hand, had more assortative tendencies. A high ratio of high-degree nodes were found linked directly to each other. The average neighborhood connectivity as a function of degree reported presence of numerous genes with both high degree and high average neighborhood connectivity, see Figure B.1. As this figure shows there was no positive linear trend, thus the network does not have assortative structure even though some components were more assortative than others.

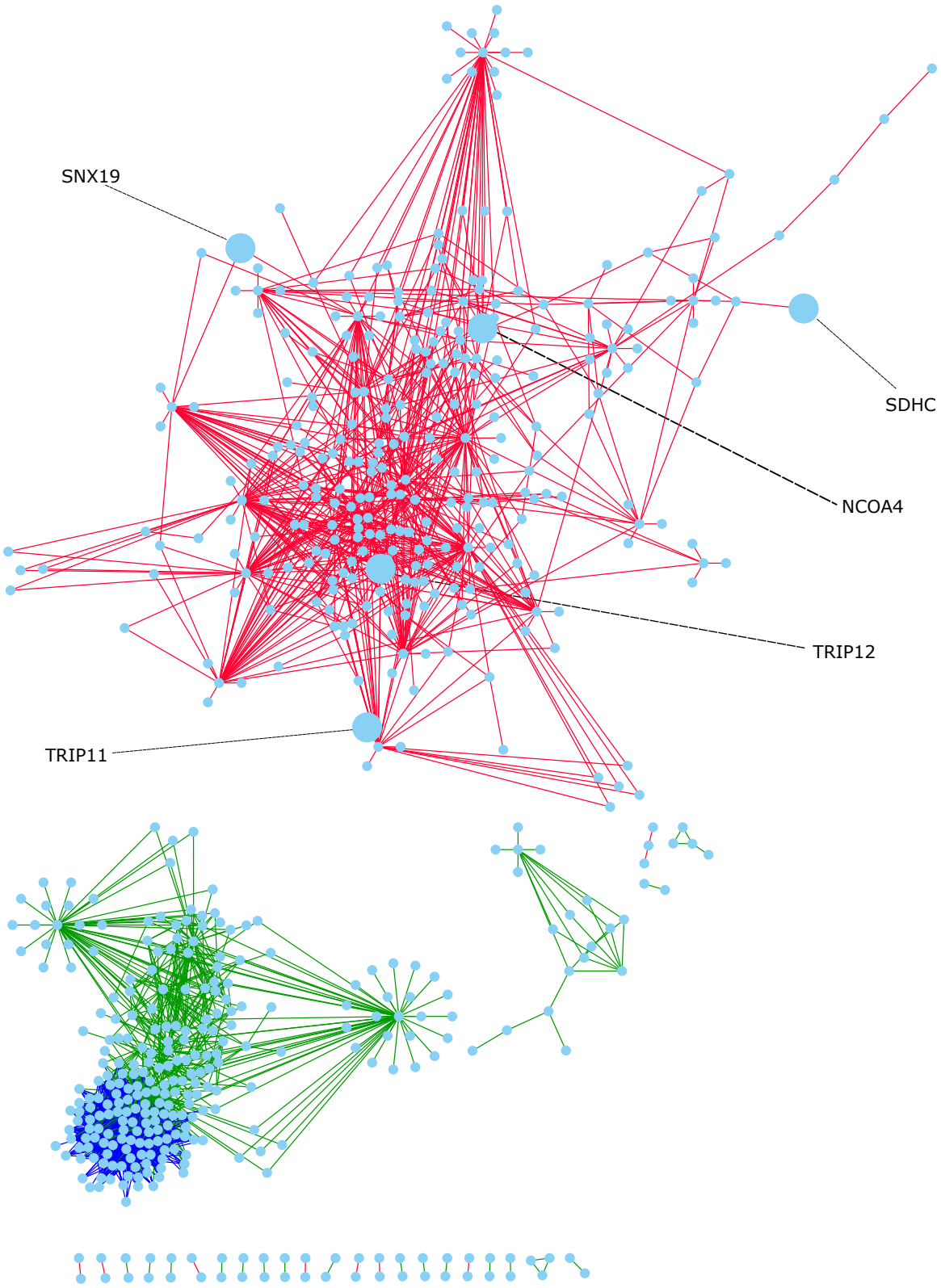


**Figure 5.3:** Differential co-expression network inferred with CSD-method implementation with weighted topological overlap (wTO) for Analysis 1, with importance level is $10^{-5}$. The network consists of 770 genes and 4831 edges. Interactions are of all types, C-scores are colored blue, S-scores are green, and D-scores are red. Thyroid cancer-associated genes are highlighted with enlarged node size and tagged by name.

### 5.2.1 GO enrichment analysis

Functional enrichment of all genes of the wTO-based network was done to get a better understanding of processes taking place in thyroid tissue that might be eriched related to the DEGs. The functional analysis result was obtained by enrichment analysis with PANTHER [110]. The wTO-network was significantly enriched for around 200 biological processes and the ones with the highest fold change are included in Table B.2. We can observe that there is indication of DEGs involved in pathways related to immunity, antibody-mediated phagocytosis, immunoglobin production and signaling, and signaling through the Fc receptor. Many of the same processes in this table are similar to those that were enriched for the standard CSD-network based on Spearman's correlation coefficient.

The Fc receptors are important regulators of the immune system. They are found on the cell surface where they interact with antibodies. Some of these bind to immunoglobulins, which causes them to induce phagocytosis. They are also involved with recognition of antibodies on the surface of abnormal cells to target and destruct cancer cells. The humoral immune system is regulated by these Fc receptors, and malignant alterations in their behaviour is associated with autoimmunity and inflammation [141]. Altered expression of Fc-receptors are speculated to play an important role in autoimmune thyroid diseases like Hashimoto's thyroiditis [142]. In this complex interplay of immunity in cancer there are many processes that require integral functionality of numerous components. Among the DEGs in Table B.2 there are many representatives of this system. There is high enrichment in processes where different Fc-receptors are involved in signaling cascades and mediating phagocytosis. Enrichment in immunoglobulin immune response are also present and highly enriched. Ultimately, these are strong indicators of striking high activation in the conduct of the immune system. An important fact about papillary thyroid carcinoma (PTC) is that it is remarkably innocuous. In PTC the high ability of the humoral immune response to detect and target the carcinogenic cells contribute to the quiescent nature of these tumors [143]. However, other types of thyroid cancer are characterized by invasive and aggressive behaviour [144]. In Table B.2 there is also a 8.98 fold enrichment of proteasomal degradation. This could be a symptom of a disrupted endoplasmatic reticulum (ER) unable to process and process proteins correctly because of chronic inflammation. Chronic ER stress is characteristic of several autoimmune thyroid diseases [145]. In thyroid tissue with these conditions there is thus reason to believe that this could mediate the impediment of the immune system and lead to a more invasive thyroid carcinoma [131, 146] .

### 5.2.2 The behaviour of network hubs

To further screen for DEGs implicated in , looking into hubs. A list of the network's hubs is given in Table B.1. The most highly connected hub genes in the wTO-network had over twice as many neighbors compared to hubs in the $\rho$-based network. There are 18 common hubs among the genes with $k > 40$ between these networks. However, the largest hubs

in the wTO-network had degrees three times as high as in the $\rho$-network. There were ten instances of genes with $k > 100$, which is extremely high given that this network has half the size and just a third more edges.

**Table 5.3:** Table of GO enriched processes for the network hubs of the wTO-network.

| GO biological process | FE |
|---|---|
| Glomerular filtration | >100 |
| Positive reg. of respiratory burst | >100 |
| Reg. of complement activation | 96.7 |
| Humoral immune response activation | 84.4 |
| Fc-gamma receptor signaling pathway involved in phagocytosis | 82.3 |
| Proteasomal ubiquitin-independent protein catabolic process | 74.1 |
| Fc-epsilon receptor signaling pathway | 67.8 |
| B cell receptor signaling pathway | 58.4 |

Table 5.3 present the enriched processes of hubs in the wTO-network, generated by GO enrichment analysis with the biological enrichment tool PANTHER [110]. Very high enrichment in glomerular filtration indicates that the thyroid tissue performs selective filtration of the blood [147], which is a normal function of thyroid tissue [16].

Many hubs in Table B.1 are genes encoding various proteins of the immunoglobulin family. These have numerous associations of conserved type. These immunoglobulin proteins thus seem to be expressed at normal levels in relation to others in both carcinogenic and normal thyroid gland tissue. Exceptions are IGKV1OR2-108 and IGLV1-50, which have 69 and 58 S-links in the network respectively. These highly connected genes encode proteins of immunoglobulins. They are listed as non-functional in the UniGene database, but their specific genetic associations suggest that the have shifted behaviour significantly in the transcriptional network. Mutations in genes encoding immunoglobin kappa (IGK) and immunoglobin lambda (IGL) are related to blood lymphocyte cancer [148]. An important observation about these immunoglobulin-encoding genes is that they are found in between a dense region of conserved type link-dominated neighborhood and a regions of specific type link-dominated neighborhood. Several are *heterogeneous* in respect to link type distribution. These hubs are co-expressing with some neighbors in a conserved way and other in a specific way. Specificity of the links indicate that co-expression patterns present in one condition is not in the other. It is thus reasonable to assume that the normal transcriptional regulation of these is significantly altered in thyroid cancer. These immunoglobulin-encoding genes could harbour physical changes themselves, causing them to interact with other cellular proteins differently. These changes could be a result of mutations in these hub genes or abnormal post-translational processing. Another possible explanation is that the transcriptional regulatory system controlling the expression levels of these genes are behaving abnormally as well. Together with the exceedingly high enrichment of pathways related to the immune

system in Table 5.3, this could suggest that potential mutations in the immunoglobulin-coding genes could be mediating pathogenic functionality of the immunoglobulin proteins driving carcinogenic processes in the thyroid tissue.

The biggest hub is the gene SP140 which is a gene encoding a component of the nuclear body (NB). This is highly expressed in cells involved with host defence [149] and is connected to 110 other genes by specific type links, indicating that it has abnormal co-expression patterns. SP140 is found in the transition between the network region dominated by specific-type (S-links) links and conserved-type (C-link) links. SP140 in homogeneously linked by S-links but many of its neighbors are heterogeneous mixes of S- and C-type dominated hubs. These are the immunoglobulin-coding hub genes mentioned above.

Other enriched processes, like complement activation and regulation of the humoral immune response are enriched to a similar high degree. Many of the hubs involved in pathways related to immunity are linked to other DEGs by specific type links. TRAC encodes a T-cell receptor alpha constant protein which is an essential specific antigen-receptor. This gene has mostly S-links, which could be an indication of mutation. Genetic alterations of TRAC are related to immunodeficiency [150]. In the transcriptomic network TRAC has 110 links to other genes and these associations change significantly from normal to carcinogenic thyroid specimens. The consequences of aberrant behaviour in these genes will thus affect the cellular immune system-related behaviour tremendously.

The enrichment in proteasomal protein catabolism can most likely be rooted in altered transcriptional pattern of the hub PSMA3. It encodes the Proteasome subunit alpha type-3, a protein that mediates ubiquitin-independent degradation of proteins in the cell [151]. Abnormal high levels of misfolded protein could be a result of the reduced functionality of the endoplasmatic reticulum in thyroid cancer cells [131]. In the network this hub has exclusively differential links, indicating that it's associations to other genes is opposite between the studied conditions. The important role in degradation of damaged and misfolded proteins could be altered, causing retardation of the cell's ability to maintain homeostasis. There are also other processes that testify of loss of homeostasis. The second process of Table 5.3 is a over hundred-fold enrichment in respiratory burst. It is a process that is related to innate immune system of self-induced apoptosis [152]. The network hub XRN2 encodes a 5'-3' exoribonuclease 2, which is a transcriptional terminator. The last hub in Table B.1 is the gene Cluster of Differentiation 53 (CD53), which encodes a cell-surface antigen that contribute to tumor cell survival [153]. All of the hubs PSMA3, XRN2 and CD53 have abnormal activity in the transcriptomic network. PSMA3 and XRN2 associate with other DEGs oppositely between normal and cancerous thyroid tissue and CD53 associate specifically to one condition. This is a strong indication that these very central players in the genetic network have a significantly altered transcriptional correlation with other genes in carcinogenic thyroid tissue and are likely to be involved in the conduct of deleterious pathways ultimately leading to a cancerous cell fate.

### 5.2.3 THCA-associated genes

Querying the DEGs from the network based on wTO as similarity measure resulted in identification of thyroid-cancer associated genes. Table 5.4 lists the thyroid-cancer associated genes identified with the wTO-network. Three additional THCA-genes were identified as a result of applying wTO as similarity measure for co-expression analysis. Total number of THCA-associated genes from this network was 5. There two genes NCOA4 and TRIP12 were identified in both the wTO-network and the standard CSD-network. In the wTO-network all five THCA-associated genes were found in the largest connected component, with exclusively differentiated type links. Both TRIP11 and TRIP12 were closely bound to the hubs PSMA3 and XRN2. The biological role of these genes in pathogenesis of thyroid carcinoma will be elaborated in A.2.

The ability of identifying biologically relevant DEGs is employed as a measure of success for the network inference method in this thesis. Here, basing the network on wTO resulted in five THCA-associated genes, three of which were not identified by the standard CSD-method. This implies that this network is biologically meaningful and that it is able to highlight genes that are significantly altered in carcinogenic thyroid gland tissue compared to normal tissue. It is anyhow substantially less successful compared to the CSD-network, as it only identified 5 THCA-associated genes which under 25%. An interesting fact though, is that all of these were different. This indicates that transforming the correlation to weighted topological overlap changes the content of genetic interactions that will be included in the inferred network.

**Table 5.4:** Table of thyroid cancer-associated genes uniquely identified from functional annotation analysis of the wTO-network, sorted by degree.

| Gene symbol | Gene name | $k$ | $t_{\in(C,S,D)}$ |
|---|---|---|---|
| TRIP12 | Thyroid hormone receptor interactor 12 | 11 | D |
| NCOA4 | nuclear receptor coactivator 4 | 9 | D |
| SDHC | Succinate dehydrogenase complex subunit C | 1 | D |
| SNX19 | Sorting nexin 19 | 2 | D |
| TRIP11 | Thyroid hormone receptor interactor 11 | 1 | D |

## 5.3 Expanding the CSD framework with mutual information

The main goal of this software development part was to develop an implementation and investigate the use of mutual information as an alternative similarity measure to Spearman's ranked correlation coefficient. As the mutual information is a more far-reaching similarity measure taking to account non-liner correlations it was interesting to compare results from networks constructed on the basis of this this alternative similarity measure. This section will first investigate quantitative properties of employing mutual information as similarity measure and a qualitative survey of it's influence of biological meaningfulness.

The application of mutual information as similarity measure for co-expression analysis was developed in a parallelized *R*-script, for rapid computation time performance. Given that it needed to handle large data sets of twenty thousand variables in each dimension of a two-dimensional array of expression vectors, parallelization was imperative. The program implemented in parallel successfully computed the similarity matrix from the expression data so that the CSD-network inference workflow was applicable to the output from this program. The code was run on a computer with 48 cores for both An.1 and An.2. The run-time performance was assayed for An.1 (16,728 genes) by storing elapsed computation times for several smaller sets from the complete one. Results are presented in Figure 5.4. The curve for the run-time complexity shows that the implementation of the script shows a near exponential growth in computational time depending on data set size. This makes this method challenging in terms of required computational power for large data sets.

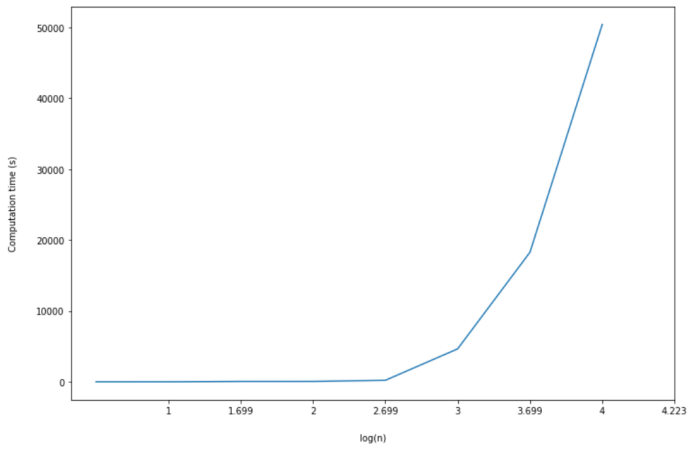


**Figure 5.4:** Extrapolated run-time complexity plot for mutual information computation of similarity matrix for differential co-expression network inference. Performance is illustrated as time elapsed as a function of $log(n)$, $n$ = number of variables (data set size).

The resulting differential gene co-expression network comparing thyroid cancer normal thyroid tissue generated by the CSD-method was substantially different in structure compared to the standard $\rho$-based network. Figure 5.5 illustrates the inferred network. Data from the MI-network is given in Table 5.5. With the same importance value used to infer the network, $p = 10^{-5}$, the MI-based network had larger size but fewer edges. The network did not contain one dominant component consisting of the majority of genes, as was observed

for both the $\rho$- and the wTO-based networks. The MI-network consisted of 532 groups of which 387 were separate pairs of two. The largest component consisted of 319 genes and several other components of intermediate size were also present.

**Table 5.5:** Table of network parameters for the networks based on mutual information (MI) and on Spearman's correlation coefficient $\rho$, both with $p = 10^{-5}$.

| $S_{ij}$ | Genes | Edges | $\langle k \rangle$ | C | Diam. | $\bar{d}$ | Power law |
|---|---|---|---|---|---|---|---|
| MI | 2106 | 2367 | 2.2 | 0.05 | 16 | 5.8 | a = 848, b = -1.9 |
| $\rho$ | 1516 | 3612 | 4.8 | 0.11 | 18 | 6.4 | a = 524, b = -1.5 |

The MI-based network had lower average degree and average clustering coefficient than the $\rho$-based network. In Figure B.2 the degree distribution for the network is shown a log-scale. The degree distribution's approximated power-law function $y = 848x^{-1.9}$ had a very high fit to the data ($R = 0.998$, $R^2 = 0.913$). A degree exponent close to -2 and a scale-free network structure are typical for metabolic networks [1], suggesting that the differential co-expression network is a reasonable representation of the underlying transcriptome profile in the thyroid cells. The strategy for correction of variance in MI-scores between expression vectors and importance level thresholding for network inference was identical to that of networks based on $\rho$ and wTO as similarity measure. The discordant structure observed here compared to the very interconnected networks based on $\rho$ and wTO are thus likely to be rooted in the mutual information as similarity measure.

### 5.3.1 The behaviour of hubs

There were only 25 DEGs in this network with $k > 20$. These are listed in B.3. Many of the hubs in this network are genes encoding immunoglobulins or genes encoding proteins involved in immune system processes. These are mostly interacting with other genes by conserved links and are homogeneous. These then are likely to be regulated normally in both conditions. Other hubs were dominated by specific type links.

The hub RPS17 encodes ribosomal protein S17-like and is connected to many other ribosomal protein-encoding genes. These are associating by S-links, indicating that their association is specific for either sick or healthy persons. Many of it's neighbors are also ribosome protein-coding but not homogeneous and mark the transition from conserved associations and specific. This suggests that there could be some general regulatory machinery acting on all of the S-linked ribosomal proteins and thus take significant control of all of these gene's expression levels. The hub USP34 encodes a peptidase that has an important role in port-translational protein modification. Mutations in this gene has a documented correlation to cancer progression and cancer cell survival [154], this gene has exclusively S-links. Another S-link dominated hub is the gene MFAP3, which encodes a microfibril-associated protein important for regulation of several other genes through phosphorylation of signal proteins such as EGFR and ERK2 which upon this regulation mediate cancer metastasis and poor

clinical outcome [155]. Aberrant behaviour in these hubs is thus likely to cause changes in the expression levels of other genes, as is seen in the high number of other genes these correlate with. Specific links report that going from one condition to the other results in a sudden correlation between thees genes.

The specific type-links between these hubs clearly demonstrate that the associations between ribosomal proteins, port-translational processor proteins and signal kinases are correlation with many other in a condition-specific context. These may be subject to abnormal translational regulation that only occurs in one condition or they themselves could be regulating other abnormally. This condition-specificity is rooted in the differences between healthy and thyroid cancer-patients, and is thus likely to represent one of the manifestations of malignant associations between genes in the transcriptomic network.

### 5.3.2 THCA-associated genes

GO-enrichment analysis of the entire gene set resulted in identification of eleven new thyroid cancer associated-genes, listed in table 5.6. Of the 18 genes from this network resulting from functional annotation analysis identified thyroid cancer-associated genes, the ten genes in this table were unique to the MI-based network. All of these genes had low degrees and either associated with their neighbors by specific or differentiated type links. The transcriptional pattern of these genes thus correlated with their neighbors' in a condition-specific manner or associated with reversed sign of correlation.

GO-enrichment of the entire network resulted in enrichment of similar biological processes as those enriched for the networks based on $\rho$ and wTO. GO-enrichment analysis of hubs did not result in any significantly enriched processes. The results from the functional annotation analysis with highest fold enrichment are listed in Table 5.7. Processes involved in the immune system were highly enriched. From the functional enrichment it can be observed that the investigated thyroid transcriptome patterns have abnormal increase in activity related to antigen processing, T-cell selection, and humoral immune response regulation.

**Table 5.6:** Table of thyroid cancer-associated genes uniquely identified from functional annotation analysis of the DEGs from MI-network.

| Gene symbol | Gene name | k | $t_{\in(C,S,D)}$ |
|---|---|---|---|
| FN1 | fibronectin 1 | 8 | S |
| CCL5 | C-C motif chemokine ligand 5 | 3 | C |
| ITGB2 | Integrin subunit beta 2 | 3 | C |
| ITPR1 | inositol 1,4,5-triphosphate receptor 1 | 3 | D |
| TPO | Thyroid peroxidase | 3 | D |
| ARRB2 | Arrestin beta 2 | 1 | S |
| DIO1 | iodothyronine deiodinase 1 | 1 | D |
| GSTP1 | glutathione S-transferase pi 1 | 1 | D |
| ITGA1 | Integrin subunit alpha 1 | 1 | S |
| ITGA3 | integrin subunit alpha 3 | 2 | D |
| HLA-DQA1 | major histocompatibility complex, class II, DQ alpha 1 | 1 | S |
| LGALS3 | galectin 3 | 2 | S |
| MLH1 | MutL homolog 1 | 1 | D |
| NCOA4 | nuclear receptor coactivator 4 | 1 | D |
| PLAU | Plasminogen activator, urokinase | 1 | S |
| PRKAR1A | Protein kinase type I reg. subunit alpha | 2 | D |
| S100A10 | S100 calcium binding protein A10 | 1 | S |
| SDHB | Succinate dehydrogenase complex subunit B | 2 | D |
| SDHC | succinate dehydrogenase complex subunit C | 2 | D |
| SLC26A4 | solute carrier family 26 member 4 | 1 | C |
| SNX19 | Sorting nexin 19 | 1 | D |
| TCF12 | Transcription factor 12 | 2 | S |
| TG | thyroglobulin | 1 | D |
| TRIP11 | Thyroid hormone receptor interactor 11 | 2 | S |
| TRIP12 | thyroid hormone receptor interactor 12 | 1 | S |

**Table 5.7:** GO functional enrichment analysis results for the CSD-network based on mutual information as similarity measure. The importance level for the network is $10^{-5}$. Enriched processes are sorted by fold enrichment.

| GO biological process | FE |
|---|---|
| SRP-dependent cotranslational protein targeting to membrane | 8.5 |
| Positive thymic T cell selection | 8.1 |
| mRNA catabolic process, nonsense-mediated decay | 7.5 |
| Translational initiation | 7.5 |
| Pos. reg. of interleukin-2 biosynthetic process | 6.9 |
| Reg. of humoral immune response | 6.5 |
| Complement activation, classical pathway | 6.4 |
| Fc-gamma receptor signaling pathway, phagocytosis | 6.3 |
| Reg. of antigen processing and presentation | 5.8 |

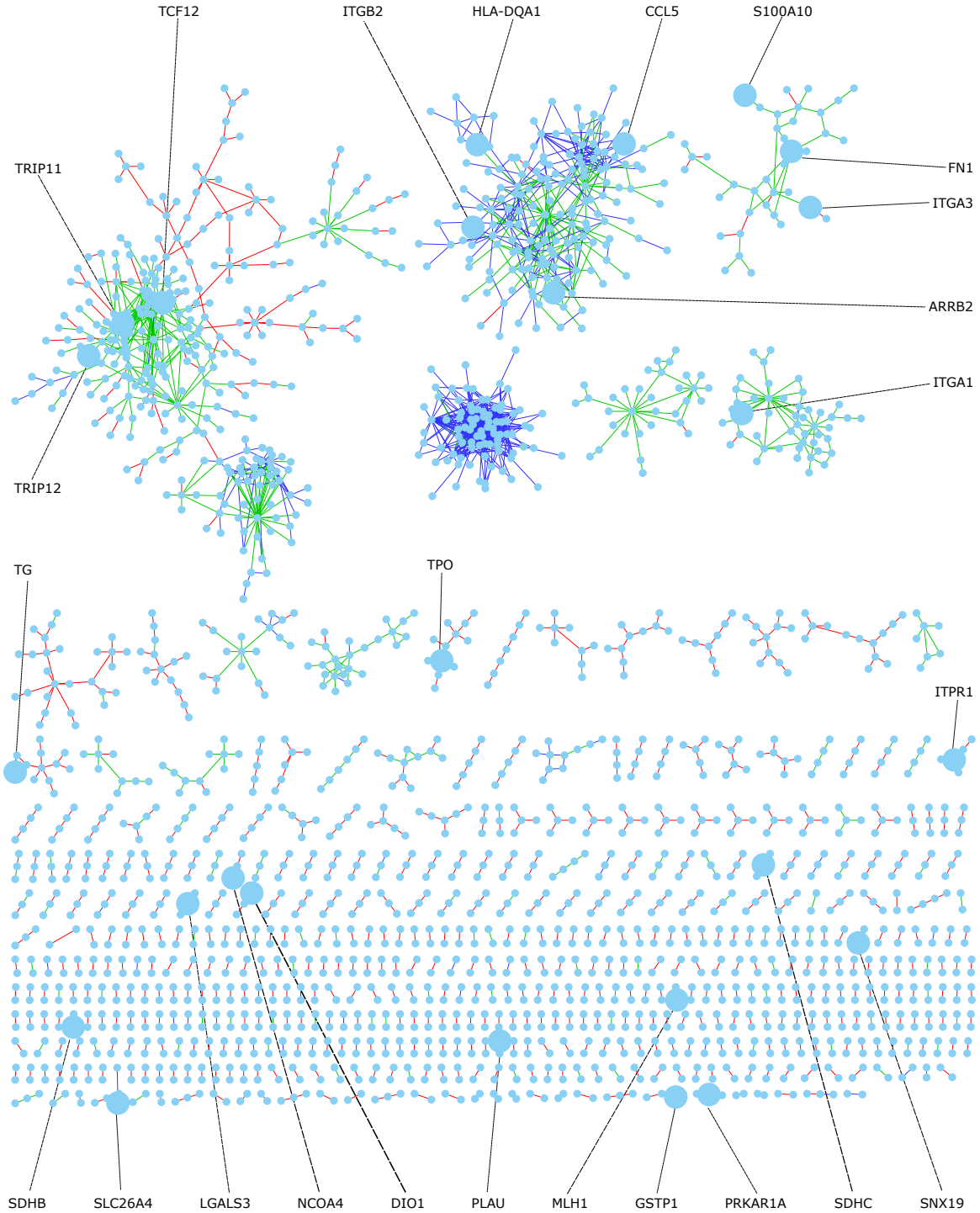


**Figure 5.5:** Differential co-expression network inferred with CSD-method implementation with foundation in mutual information as similarity measure. Importance level is $10^{-5}$. Interactions are of all types, C-scores are colored blue, S-scores are green, and D-scores are red. Thyroid-cancer genes are tagged by name and highlighted with enlarged node size.

## 5.4 Comparison of co-expression measures

The following section will present results from comparison of the different networks inferred with the CSD-framework based on four similarity measures. An illustration of how these alternatives differ is given in Figure 3.2.7. The results demonstrates the effect various similarity measures forming the foundation for co-expression analysis have on the inferred networks. All networks have been inferred from the identical data set of thyroid cancer expression data from TCGA and healthy control from GTEX measured in thyroid glad tissue.

Results for networks constructed with four different similarity measures are basis for the CSD-framework for differential co-expression network inference. Similarity measures are used to calculate the values for the correlation matrix, from which the network adjacency matrix is inferred with the CSD-framework. These four strategies are abbreviated in the following manner:

- CSD: Spearman's rank correlation coefficient. As original implementation of CSD-framework
- CSD -VAR : Spearman's rank correlation coefficient without variance correction.
- wTO : Weighted topological overlap. Computed by topological overlap matrix transform of correlation matrix based on Spearman's rank correlation coefficient, soft thresholded using $\beta = 5$.
- MI : Mutual information inferred with *parmigene*, where the variance of MI estimation per gene-pair of an association is based on 20 replicates.

Comparison of Spearman correlation and mutual information: In both of these measures, in which two gene's expression values $X$ and $Y$ are random variables, each of the similarity scores contains a point-wise measure of the distance of the two random variables from independence. They are both expressed as a distance between the random variables' joint probability mass functions (PMF) $p(x)p(y)$ from the product of the variables' marginal PMF's. For the correlation measure, this distance is in the form of ranked gene expression levels and for the MI difference in the form of logarithms. The similarity measures differ in that the correlation create a weighted sum of the product of the two random variables, while the MI measure what independence between the variables - or, rather, a lack thereof - does to their joint probability.

### 5.4.1 Network construction

For both versions of transcriptomic data sets preprocessed in different ways, each of them were used to generate differential co-expression networks with the CSD-method with different cut-off values. The table of page 100 shows the network parameters for each of them. Both data sets consisted of 504 thyroid cancer samples and 399 normal controls. Analysis 1

is the data set with identical pre-processing of both the thyroid cancer and healthy control transcriptomic data with strict filtration. This set consists of 16,728 genes. In Analysis 2 the thyroid cancer data had been less strictly filtered, and contained 20,657 genes.

Genes and edges correspond to the number of genes and edges in the networks respectively, $\langle k \rangle$ is the average degree, $C_v$ is the average clustering coefficient. The importance level $p$ is the common importance level used to estimate threshold values $X_p^{C,S,D}$ for each of the scores in order to map them in a comparable scale [3]. The table also lists the network diameter (Diam.), density, average shortest path, and the variables of the degree distribution fitted power law, who's expression is given by $y = ax^b$. The variables $a$ and $b$ are listed, together with the $R$ and $R^2$ values for the evaluated power law fit to the data.

| Analysis 1 | p | Genes | Edges | $\langle k \rangle$ | $C_v$ | Diam. | Density | Path | Power law |
|---|---|---|---|---|---|---|---|---|---|
| CSD | $10^{-4}$ | 4700 | 24301 | 10.341 | 0.116 | 13 | 0.002 | 4.788 | a = 3146.7. b = -1.590. R = 0.978. $R^2 = 0.926$ |
| | $10^{-4.5}$ | 2695 | 8292 | 6.154 | 0.095 | 22 | 0.002 | 6.363 | a = 1525.1. b = -1.667. R = 0.998. $R^2 = 0.894$ |
| | $10^{-5}$ | 1313 | 2442 | 3.72 | 0.074 | 23 | 0.003 | 3.72 | a = 646.39. b = -1.716. R = 0.997. $R^2 = 0.927$ |
| | $10^{-5.5}$ | 587 | 804 | 2.739 | 0.085 | 14 | 0.005 | 5.221 | a = 277.7. b = -1.737 . R = 0.999. $R^2 = 0.913$ |
| CSD-VAR | $10^{-4}$ | 5118 | 26327 | 10.288 | 0.106 | 16 | 0.002 | 4.997 | a = 3373.2. b = -1.584. R = 0.977. $R^2 = 0.932$ |
| | $10^{-4.5}$ | 2945 | 9031 | 6.133 | 0.093 | 20 | 0.002 | 6.383 | a = 1878.1. b = -1.799. R = 0.985. $R^2 = 0.928$ |
| | $10^{-5}$ | 1448 | 2725 | 3.764 | 0.080 | 19 | 0.003 | 5.964 | a = 704.51. b = -1.720. R = 0.998. $R^2 = 0.922$ |
| | $10^{-5.5}$ | 647 | 806 | 2.491 | 0.064 | 16 | 0.004 | 5.335 | a = 315.88. b = -1.835. R = 1. $R^2 = 0.895$ |
| wTO | $10^{-4}$ | 2093 | 24650 | 23.555 | 0.166 | 14 | 0.011 | 3.668 | a = 294.18. b = -1.062. R = 0.996. $R^2 = 0.764$ |
| | $10^{-4.5}$ | 1207 | 9729 | 16.121 | 0.134 | 10 | 0.013 | 3.237 | a = 169.61. b = -1.026. R = 0.998. $R^2 = 0.8$ |
| | $10^{-5}$ | 608 | 2659 | 8.747 | 0.155 | 9 | 0.014 | 3.195 | a = 142.73. b = -1.174. R = 0.997. $R^2 = 0.861$ |
| | $10^{-5.5}$ | 325 | 812 | 4.997 | 0.184 | 18 | 0.015 | 3.570 | a = 86.579. b = -1.239. R = 0.990. $R^2 = 0.841$ |
| MI | $10^{-4}$ | 8683 | 22653 | 5.218 | 0.051 | 17 | 0.001 | 6.146 | a = 5783.8. b = -1.825. R = 0.980. $R^2 = 0.927$ |
| | $10^{-4.5}$ | 4623 | 7517 | 3.252 | 0.050 | 23 | 0.001 | 8.781 | a = 2371.0. b = -1.882. R = 0.998. $R^2 = 0.936$ |
| | $10^{-5}$ | 1807 | 2775 | 3.071 | 0.065 | 26 | 0.002 | 8.604 | a = 721.71. b = -1.766. R = 1. $R^2 = 0.902$ |
| | $10^{-5.5}$ | 356 | 303 | 1.702 | 0.035 | 7 | 0.005 | 2.883 | a = 202.39. b = -2.153. R = 0.999. $R^2 = 0.921$ |

| Analysis 2 | p | Genes | Edges | $\langle k \rangle$ | $C_v$ | Diam. | Density | Path | Power law |
|---|---|---|---|---|---|---|---|---|---|
| CSD | $10^{-4}$ | 6593 | 32569 | 9.880 | 0.101 | 18 | 0.001 | 4.921 | a = 4170.9. b = -1.589. R = 0.985. $R^2 = 0.920$ |
| | $10^{-4.5}$ | 3535 | 11709 | 6.625 | 0.104 | 16 | 0.002 | 6.065 | a = 1343.2. b = -1.511. R = 0.998. $R^2 = 0.912$ |
| | $10^{-5}$ | 1516 | 3612 | 4.765 | 0.105 | 18 | 0.003 | 6.391 | a = 524.47. b = -1.523 . R = 0.998. $R^2 = 0.887$ |
| | $10^{-5.5}$ | 694 | 1213 | 3.496 | 0.100 | 23 | 0.005 | 6.970 | a = 330.1. b = -1.636. R = 0.996. $R^2 = 0.937$ |
| CSD-VAR | $10^{-4}$ | 6618 | 38881 | 11.750 | 0.124 | 13 | 0.002 | 4.769 | a = 3950.5. b = -1.526. R = 0.985. $R^2 = 0.921$ |
| | $10^{-4.5}$ | 3621 | 13359 | 7.379 | 0.114 | 21 | 0.002 | 6.258 | a = 1760.5. b = -1.559. R = 0.995. $R^2 = 0.931$ |
| | $10^{-5}$ | 1594 | 4194 | 5.262 | 0.123 | 15 | 0.003 | 5.858 | a = 686.24. b = -1.569. R = 1.0. $R^2 = 0.924$ |
| | $10^{-5.5}$ | 719 | 1360 | 3.783 | 0.125 | 13 | 0.005 | 3.783 | a = 344.7. b = -1.630. R = 0.998. $R^2 = 0.943$ |
| wTO | $10^{-4}$ | 2824 | 27498 | 19.475 | 0.172 | 12 | 0.007 | 3.916 | a = 433.13. b = -1.121. R = 0.996. $R^2 = 0.698$ |
| | $10^{-4.5}$ | 1419 | 14248 | 20.082 | 0.181 | 16 | 0.014 | 3.979 | a = 130.06. b = -0.938. R = 0.968. $R^2 = 0.576$ |
| | $10^{-5}$ | 770 | 4831 | 12.548 | 0.274 | 11 | 0.016 | 3.547 | a = 190.01. b = -1.169. R = 0.993. $R^2 = 0.844$ |
| | $10^{-5.5}$ | 430 | 1241 | 5.772 | 0.333 | 10 | 0.013 | 3.385 | a = 100.18. b = -1.233. R = 0.977. $R^2 = 0.774$ |
| MI | $10^{-4}$ | 5406 | 10405 | 3.849 | 0.040 | 26 | 0.001 | 8.264 | a = 1381.2. b = -1.632. R = 0.995. $R^2 = 0.852$ |
| | $10^{-4.5}$ | 2405 | 3725 | 2.986 | 0.045 | 31 | 0.001 | 9.424 | a = 485.03. b = –1.542. R = 0.985. $R^2 = 0.866$ |
| | $10^{-5}$ | 980 | 1156 | 2.359 | 0.057 | 6 | 0.002 | 2.348 | a = 240.53. b = -1.591. R = 0.973. $R^2 = 0.865$ |
| | $10^{-5.5}$ | 396 | 382 | 1.929 | 0.047 | 8 | 0.005 | 3.048 | a = 139.86. b = -1.750. R=0.991. $R^2 = 0.889$ |

Figure 5.6 illustrates the standard CSD-network inferred with the data set of Analysis 1. The thyroid cancer expression data set was preprocessed identically as the healthy thyroid expression-data from GTEx had already had been pre-processed. The importance value of the network was $10^{-5}$.

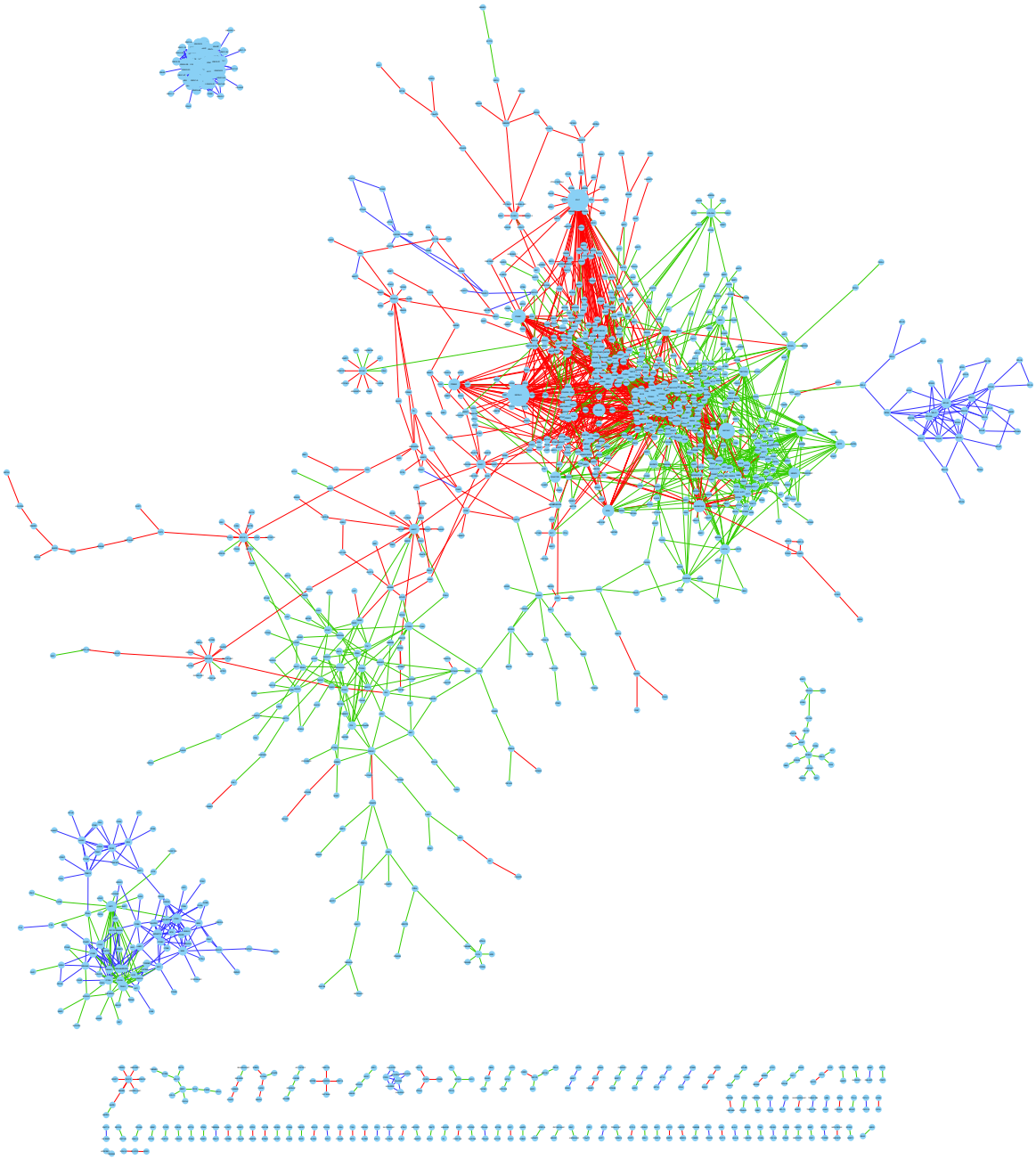

**Figure 5.6:** Differential co-expression network inferred with original CSD-implementation based Spearman's rank correlation coefficient. Interactions are of all types, C-scores are colored blue, S-scores are green, and D-scores are red. Node degree distribution is included in the left corner of the figure.

Figure 5.7 presents the network inferred with the CSD-framework with the alternative equations for the gene association-scores conserved (C), specific (S), and differentiated (D). Here these gene-pair relation scores are not corrected for (divided by) the variation in the correlation value for the two genes in the pair. The network presented is inferred with an importance value of $10^{-5}$. The node degree distribution is presented in B.3, where a red line indicates the approximated power-law function to the empirical data points.

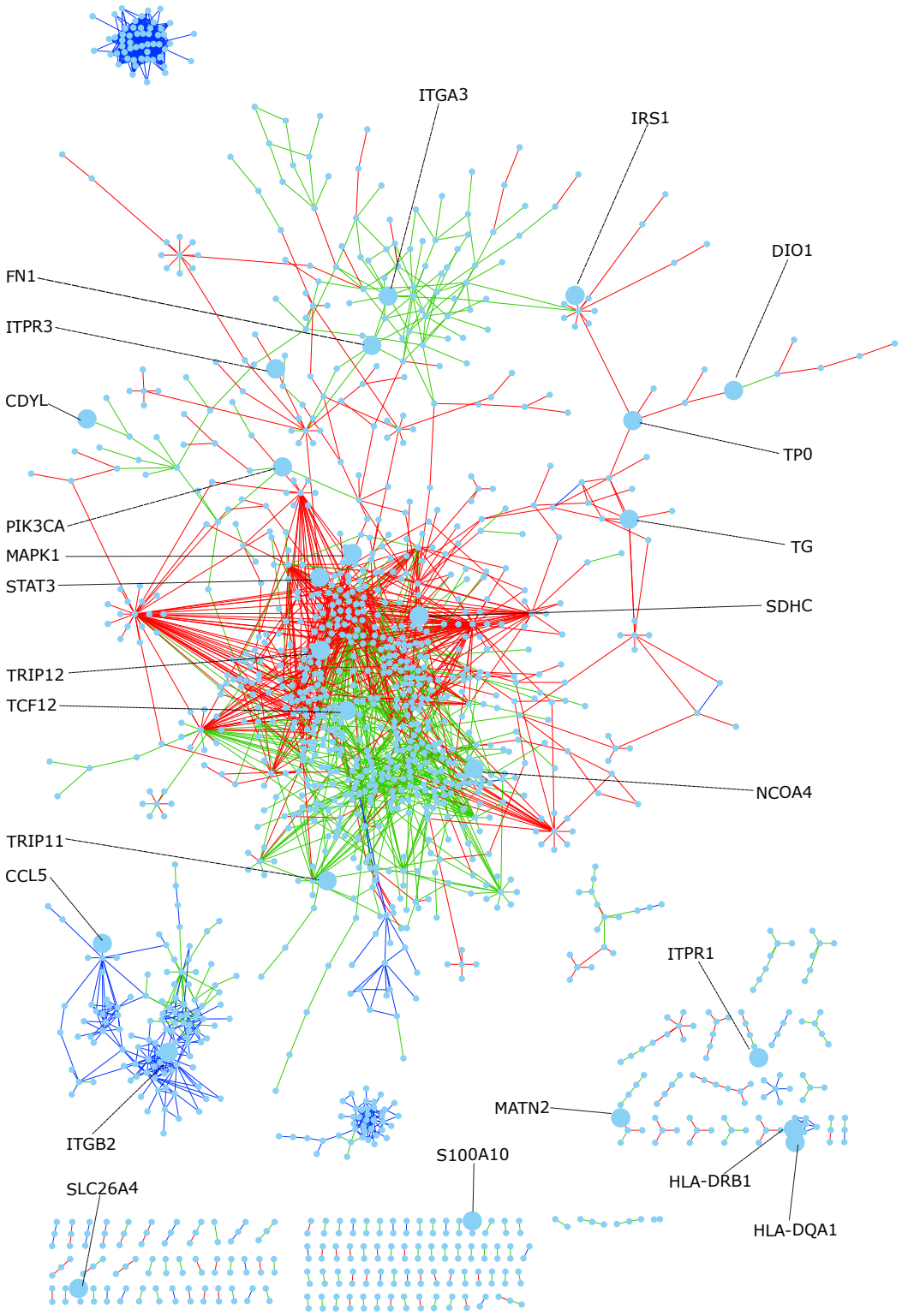


**Figure 5.7:** For network 1: Differential co-expression network inferred with CSD-method implementation with Spearman's rank correlation coefficient without variance correction (CSD-VAR). Interactions are of all types, C-scores are colored blue, S-scores are green, and D-scores are red. Node degree distribution is included in the left corner of the figure. Thyroid cancer-associated genes are highlighted by enlarged node size and tagged by gene name.

### 5.4.2 Sample size robustness

The network robustness analysis aimed to elucidate differences in constructed CSD-networks originating from the similarity measure applied as the elementary step of the differential gene co-expression (DGCE) analysis. The similarity measure $S_{ij}$ quantifies the degree of similarity in expression values between two genes in the input transcriptomic data files, here the network robustness dependent of similarity measures was tested. Four alternative similarity measures were employed and their effects on the networks were tested by robustness analysis, as explained in detail in Chapter 3.2.7. The aim of this section in to test the fidelity of a low sample-size based DGCE network compared to the network based on all samples, and use this information to compare robustness of the four similarity measures. The size of the simulated sample size, constructed with random selection of samples from the whole data set, increases along the x-axis. Three random sub-sets were constructed from the gene expression data set of sizes 50 (samples), 100, and 200, for which DGCE networks were constructed for each of the four similarity measures. The maximum number of samples were 504 samples from the thyroid cancer data set and 399 healthy persons, this set acted as baseline for assessing the sub-samples' performance. The results presented in Figure 5.8 illustrate how a small sample size of transcriptomic data set affects the quality of the inferred CSD-network.

The four similarity measures under investigation is the Spearman's ranked correlation coefficient (CSD), the Spearman's ranked correlation coefficient without variance correction (CSD-VAR), weighted topological overlap (wTO), and mutual information. The results from robustness analysis of each of them is divided into four sub-plots. Each of the four sub-plots contains either all (a) or one (b-d) type of differential gene association, of the link types conserved (C), differentiated (D) and specific (S). For each similarity measure, there are DGCE networks constructed on the basis of 50, 100 and 200 samples, and the overlap in terms of identical gene co-expression relationships (links) among them is quantified with the overlap coefficient. For the standard differential co-expression networks based on Spearman's correlation coefficient (CSD), the overlap in link content between the full set with the replicas based on 50, 100 and 200 samples are indicated with blue crosses in Fig. 5.8. The same measurements done for CSD-VAR and it's three sub-sample-based DGCE networks are indicated with red crosses. The analogous for wTO is marked green, and for MI the markers are yellow.

The y-axis represents the overlap-coefficient range from 0 to 1. A cross in the plot corresponds to the overlap coefficient in link content between the DGCE network based of the size indicated on the x-axis compared to the full set. The placement of the cross in the vertical direction is proportional to the number of common inferred links in the network based on fewer samples compared to the network based on the maximum number of samples. In this way the markers denote how fast the performance of each similarity measure depends on sample size in the data. The overlap coefficient marker for the whole sets is indicated in Fig. 5.8 by a black cross with overlap coefficients of 1. This cross is common for all

similarity measures, because it is the reference for all of their sub-sets individually and thus has full overlap with itself (overlap coefficient = 1).

The robustness assessment in Fig. 5.8 shows that for wTO, networks based on decreasing number of samples had overall higher fidelity to the full set, in terms of similar link content, than any other similarity measure. From plot a) is is clear that networks based on wTO has the highest degree of overlap in inferred gene-pair associations of all types between the full set and all sub-sample networks ranging from size 200 to 50. Here it can also be observed that CSD-VAR performs better than conventional CSD. Mutual information performs the worst, with the most rapid decline in overlap coefficient for smaller sample sizes of data compared to the full set.

In plot b), it is apparent that the conserved type link is more robust to small sample size than the other association scores, specific in plot c) and differentiated in plot d) respectively. Additionally, wTO is the similarity measure showing the highest level of resilience to low sample size. However, CSD-VAR comes close in terms of performance. It should be noted that the very low overlap coefficient between sub-sample of size 200 for CSD is an anomaly. This could be caused by two alternative events. Either it can be caused by a random enrichment of expression values of low value from the included samples of the 200-sample set, which could be excluded in the final network inference step of the CSD framework. Alternatively, it could be caused by an error in the program pipeline re-shuffling the random selection of samples chosen to constitute this sub-set. Anyhow, the two other sub-samples for CSD support the observation that conventional CSD performs worse than both wTO and CSD-VAR in terms of C-type association overlap. Standard CSD performs better than MI here also.

The plot in c) shows that S-type links are very sensitive to sample size. It is immediately obvious that specific type links are less robust than conserved links, because all overlap coefficients for every alternative similarity measure declines rapidly. This plot illustrates once again that the similarity in inferred significant gene associations in networks based on wTO are the least sensitive to decreasing sample size. The difference between CSD and CSD-VAR seems to be less pronounced, while MI is still the least resilient similarity measure. The last plot, plot d), shows that the general lack of robustness is even more pronounced in associations of differentiated type. Here, all similarity measures performs relatively similar. The first observation of wTO outperformed by another similarity measures is found for sub-samples of size 100 in this plot, where both standard CSD and CSD-VAR is somewhat more robust than wTO. MI performs the worst, differentiated associations in CSD networks based on MI are especially dependent of sample size, this MI as similarity measure suffers from limited robustness in it's applicability to data sets.

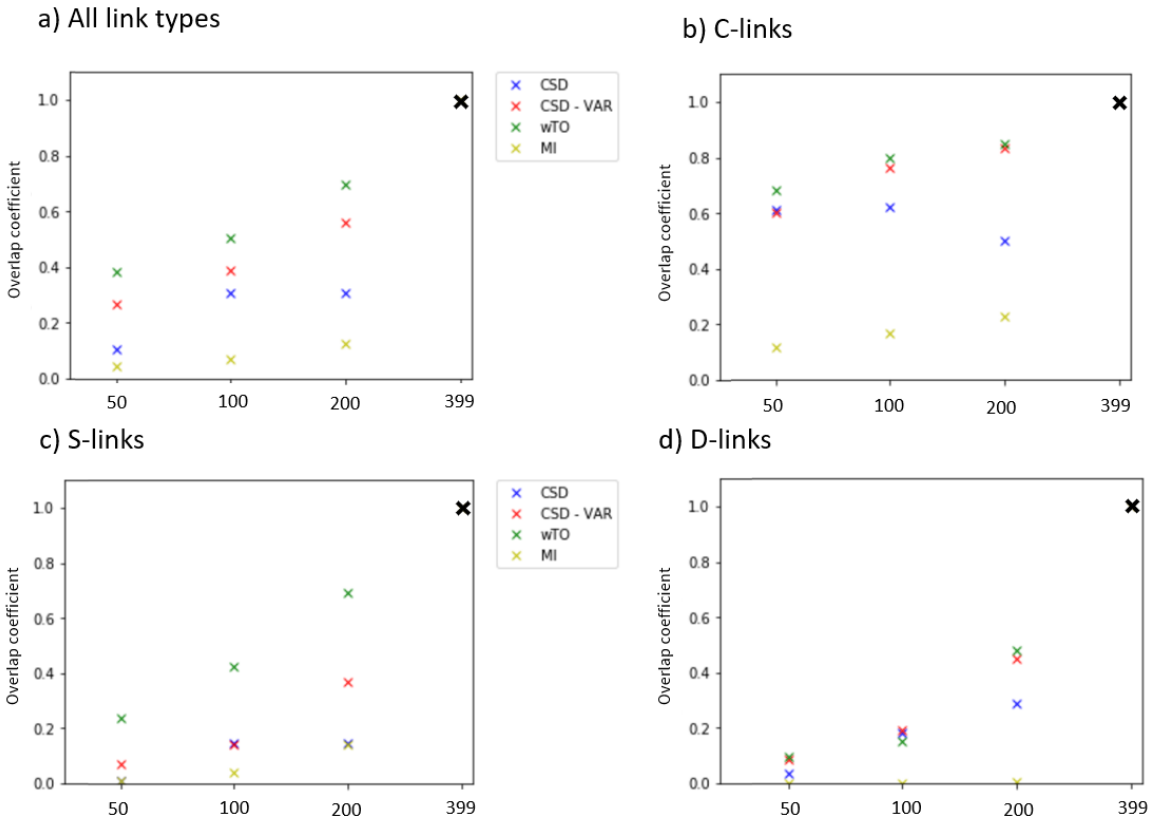


**Figure 5.8:** Robustness analysis plot for the four different similarity measures applied for differential co-expression analysis for data sets of a fabricated decreasing sample size. The overlap coefficient quantifies the degree of overlap between inferred gene associations in the networks from the sub-sample and the full set. The four different similarity measures compared are the Spearman correlation coefficient, the Spearman correlation coefficient without variance correction, weighted topological overlap and mutual information. a) Associations are of all types, C-, S-, and D. b) Associations are of conserved type (C-links). c) Associations are of specific type (S-links). d) Associations are of differentiated type (D-links).

### 5.4.3 Homogeneity

The following figures will show homogeneity results for networks constructed with four different similarity measures as basis for the CSD framework for differential co-expression network inference. All of them are constructed on the basis of CSD networks inferred with importance threshold $p = 10^{-5}$. For all of the four following figures the following applies: Red bars correspond to the median of H, and the green squares denote the mean H. The top and bottom ends of the boxes represent first and third quartile (25th percentile and 75th percentile) respectively. The ends of the whiskers represent the minimum and maximum values of H for the given degree.

Figure 5.9 illustrates the homogeneity sorted by node degree in a box-plot for the standard CSD network based on Spearman's ranked correlation coefficient. Here all low-degree nodes with $k \geq 10$ and higher-degree nodes with $k \leq 30$ have very high degree of homogeneity in distribution of association types. Intermediate degree-nodes are more heterogeneous. Some outliers are found for degrees $k = 24, k = 25, k = 35, k = 43$, where homogeneity is

substantially lower.

Figure 5.10 is the analogous box-plot for the CSD network based on Spearman's correlation without correlation-variance correction. It is sorted by node degree, which increases along the x-axis. This plot shows that there is a high average degree of homogeneity in gene association type distribution for nodes of similar degree. For intermediate node degrees, such as $k = 22, 30, 43$ there are more degree of association type mixing. This indicates that these nodes could represent genes that are involved in transcriptional regulation mechanisms across several pathways and thus have several associations of different types to their neighbor genes.

Figure 5.11 is a homogeneity box-plot sorted by node degree for the CSD-network based on weighted topological overlap binned by node degree. Here the level of homogeneity for nodes of similar degree appears to be strongly related and follow an asymptotic line towards an H-score between 0.9 and 1.0. The length and whiskers of the boxes are very short for association type distributions for node degrees above 40. This indicates that there are less outliers, and an approximately uniform level of homogeneity for the highest-degree nodes in the wTO-network. This well reflects the high degree of segregation between network components characterized by one single association type in the wTO-based CSD network, as clearly illustrated in Fig. 5.3.

The last homogeneity plot in Figure 5.12, is the box-plot of homogeneity scores sorted by node degree for the MI-based CSD network. This box-plot report of a high degree of heterogeneity in association type distribution for nodes of identical degree among the less highly connected nodes. Here, there are numerous red bars (denoting the medians) which are low. Vertically long boxes also demonstrate a larger range between first and third quartiles of the values for each node degree. In this network nodes of degree $k > 20$ are substantially more homogeneous than those with degree $k \leq 20$. Above this degree nodes are hubs and virtually homogeneous.

Venn-diagram showing the relative quantities of genes involved in each type of interaction is presented in Figure 5.13. This figure contains four Venn-diagrams containing information about the relative quantities of genes involved in each type of interaction as inferred by the four alternative similarity measures.

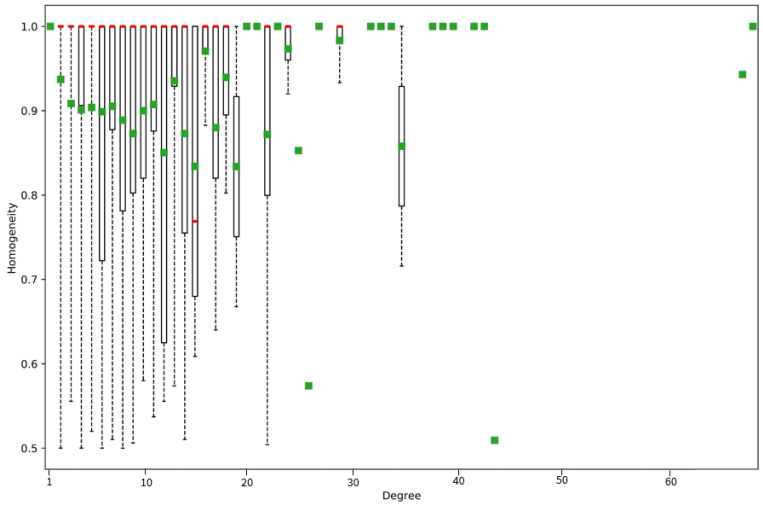


**Figure 5.9:** Plot showing homogeneity sorted by node degree for network constructed with Spearman's rank correlation coefficient (CSD) as similarity measure. The network is inferred from the data-set from Analysis 1 with an importance value $p = 10^{-5}$.

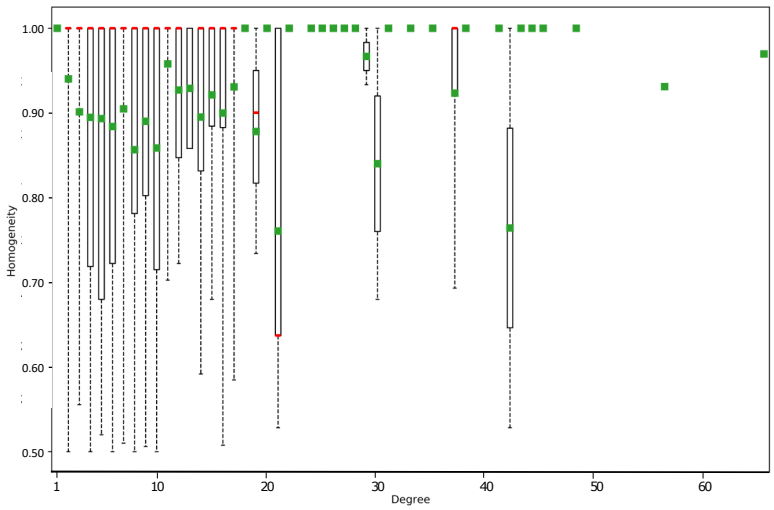


**Figure 5.10:** Plot showing homogeneity sorted by node degree for network constructed with Spearman's rank correlation coefficient without variance correction (CSD-VAR) as similarity measure. The network is inferred from the data-set from Analysis 1 with an importance value $p = 10^{-5}$.

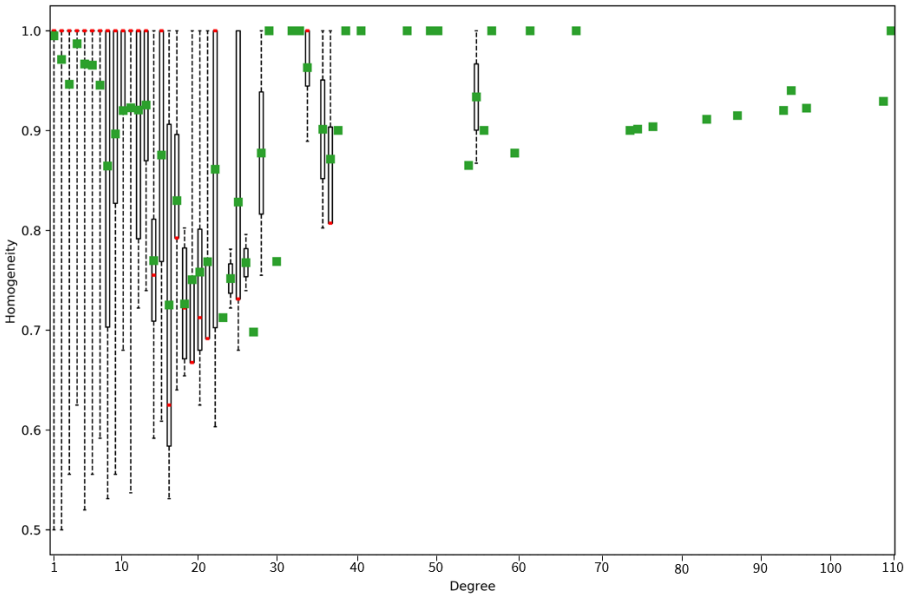


**Figure 5.11:** Plot showing homogeneity sorted by node degree for network constructed with weighted topological overlap (wTO) as similarity measure.The network is inferred from the data-set from Analysis 1 with an importance value $p = 10^{-5}$.

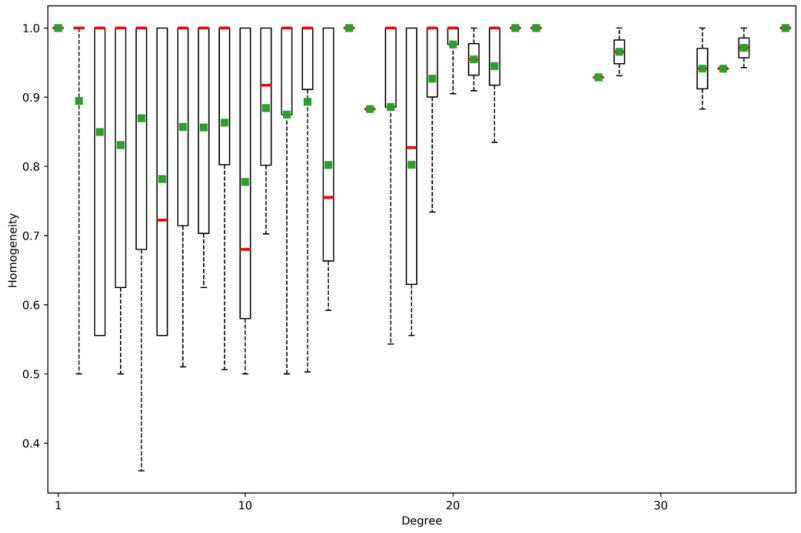


**Figure 5.12:** Plot showing homogeneity sorted by node degree for network constructed with mutual information (MI) as similarity measure. The network is inferred from the data-set from Analysis 1 with an importance value $p = 10^{-5}$.

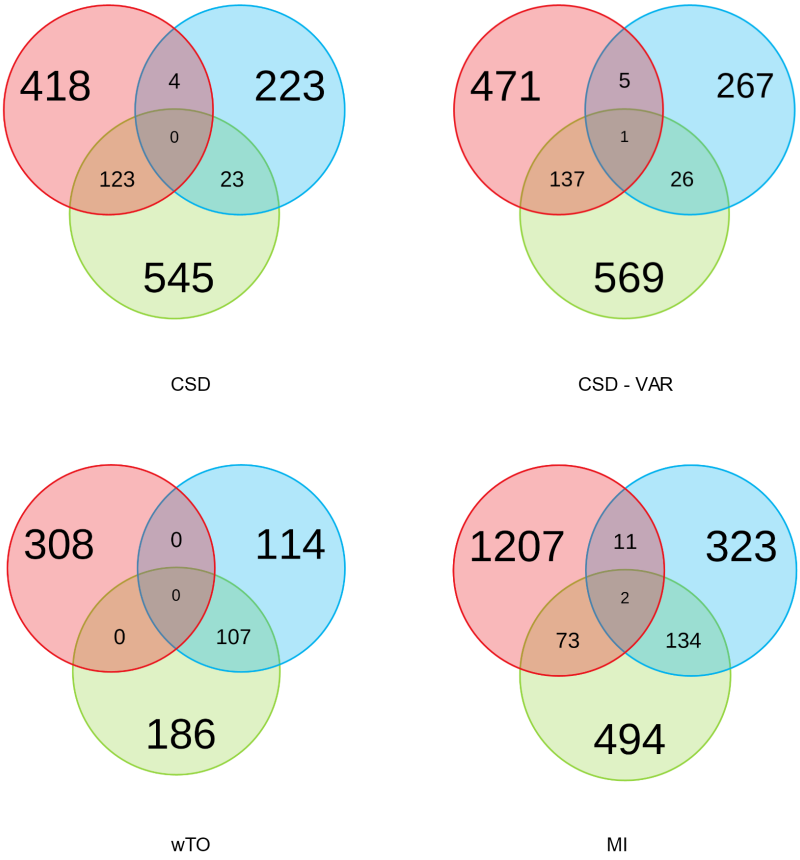


**Figure 5.13:** Venn diagram showing the relative quantities of genes involved in each type of interaction. Red circles contain the number of differentiated links, blue circles the conserved links and the green circles the number of specific type links. The networks are all inferred from the data-set from Analysis 1 with an importance value $p = 10^{-5}$.

### 5.4.4 Identification of disease genes

The similarity measures were finally assessed by their ability to infer DEGs with biological relevance. This was quantified by the presence of thyroid cancer-associated genes among the DEGs in the inferred networks. For each network based on the three new similarity measures, the number of thyroid cancer-associated present were computed and corrected for different network sizes. This represents the relative performance in this assessment. In addition, it was interesting to examine whether any similarity measure was able to *exclusively* identify any differential co-expression patterns with these disease-genes, hence eventual unique identifications were quantified as well. The result of this final assessment is presented in Table 5.8.

**Table 5.8:** Assessment of relative ability in disease-gene identification for all three alternative similarity measures compared to the baseline CSD method

| Parameter | CSD | CSD-VAR | wTO | MI |
|---|---|---|---|---|
| Size | 1313 | 1448 | 608 | 1807 |
| THCA genes | 22 | 24 | 5 | 25 |
| Rel. performance | 0.00167 | 0.00165 | 0.00822 | 0.01384 |
| THCA genes common | - | 19 | 2 | 14 |
| THCA genes new | - | 5 | 3 | 11 |

Note however, that the focus of all these CSD networks as foundation for differential gene co-expression analysis is to identify significant correlations among *pairs* of genes, and not to identify individual DEGs. Therefore we cannot expect that all or even the most prominent DEGs for thyroid cancer to be pin-pointed by these techniques for generation of differential gene co-expression networks. Co-expression networks aim to capture more complex biological processes by characterizing gene *interactions*. This is based on the assumption that pathways of cellular life is orchestrated by the interplay of many components.

# Chapter 6

# Discussion

## 6.1 Application to thyroid cancer

The first aim of this thesis was to identify inclinations for transcriptional alteration when studying expression data from thyroid cancer patients and comparing them to those of healthy persons. As presented in the results of Chapter 4, the outcome of applying differential gene co-expression analysis on transcriptomic data provided important new insights on the transcriptional network, the most important genes and the most biologically relevant clusters of genes. The CSD framework on this data proved to be very applicable and promote identification of dysregulated pathways of relevance to the disease under study.

Thyroid cancer is a clinically heterogeneous carcinoma and the most common malignant glandular tumor in iodine-sufficient countries [18]. Thyroid cancer is a type of cancer without standard Mendelian inheritance properties, yet it has the highest relative risk of first-degree relatives [156]. This motivates an investigation of genetic interplay characteristics to this disease. Here, differential co-expression network analysis in the CSD framework was employed to study the co-expression profiles of 504 samples of thyroid tissue diagnosed with cancer against 399 normal controls. Their expression patterns were used to detect significant associations between differentially expressed genes (DEGs). These resulting DEGs showed discordant correlations in expression between the conditions, indicating several abnormalities in gene expression patterns in carcinogenic thyroid tissue. Some of the DEGs were successfully identified as disease genes by query in the Gene Association Database [5]. The transcriptomic networks resulting from the CSD analysis were scale-free and good representations of the complex cellular system. This was both an indication of validity of the methods employed in this thesis and the results were supportive of recent literature of molecular mechanisms related to thyroid cancer pathogenicity.

The primary mean of identifying significant pattern changes between DEGs was through the application of the CSD framework for differential co-expression network inference. The CSD analysis is an analysis of genetic *interactions*. It aims to elucidate different types

of changes in correlation between expression levels of genes. By highlighting significantly correlated, anti-correlated and condition-specific relationships between expression levels of genes, large regions of the networks with altered transcriptional profiles were identified. A great benefit of the CSD method is that it facilitates biological interpretation because it readily distinguishes between three different types of transcriptional correlation. The method incorporates variance correction and importance levels to ensure significance and statistical quality of the inferred relationships between genes in the network. Application of this framework on the co-expression data resulted in networks for which the degree distribution closely followed power-laws and displayed hierarchical structures. This showed that the CSD networks were good representations of complex biological systems.

A central theory of network biology is that cellular functions are organized in hierarchical structures [78]. The sought for central players and modular structures within the network is therefore reasonable because they these are likely to harbour important functions in the cell. The CSD network successfully produced networks which highlighted important genes who's association with others were likely to have large impacts on the thyroid cell transcriptional system, many such genes were hubs in the CSD network (Table 4.1). There were many homogeneous S- and D-type hubs. Hubs are important in the transcriptomic network and associated with disease when mutated or misregulated. The organizational importance of hubs in networks supports their biological significance of network architectures [80]. This is further supported by the fact that many hubs were found between denser regions, as demonstrated in Table 4.4. Some of the hubs were key regulators of cell division, motility and invasive behaviour. Mutation or misregulation in some hubs have previously been linked to thyroid carcinoma. In addition, genes involved in the immune system also appeared as hubs in the network. Increased degradation of the ER by highly-connected mannosidases may be a potential mechanism by which ER stress may perturb the entire transcriptomic system and cause parthenogenesis in cells of the thyroid. Both the hubs and the CSD networks in general were highly enriched for genes and processes descriptive of abnormalities within the system (Table 4.2). This affirms the assumptions that biologically meaningful co-expression patterns were present and successfully captured in these differential co-expression networks.

The outcomes of biological process enrichment was two-fold. Firstly, it was a potent tool for investigation of characteristic behaviours for groups of DEGs. Enrichment provided many interesting insights into the interplay of processes taking place in thyroid tissue related to thyroid cancer as reported by [131]. Enrichment made it possible to establish a high degree of quality of the inferred network.

In autoimmune thyroid diseases, the misregulation of immune components degrading the proteins in the ER of thyroid cells cause inflammation which is linked to development and progression of cancer. Alteration in regulation of ER-associated mannosidases may drive inflammation of the ER and contribute to tumor progression in thyroid gland tissue [145]. Interestingly, the difference between healthy and carcinogenic co-expression patterns were drastically different for most hubs. Their altered interaction pattern are thus likely to take part in driving aberrant cellular pathways ultimately leading to a diseased phenotype.

Actually, the ability to segregate between the two studied conditions in the network was high. This was found from investigation of network and hub association type homogeneity. This altogether supported the validity of the differential gene co-expression networks inferred with the CSD framework.

Extremely high fold enrichment was observed for processes of the endoplasmatic reticulum (ER) and the immune system. High fold enrichment of processes related to the (ER) could be attributed to the functional importance of the ER in secretory cells, as extended ER is a typical trait of thyroid cells secreting thyroid hormones. But the extreme fold enrichment of the process on the basis of DEGs can also report presence of abnormal chronic stress in the ER and loss of immune system integrity, previously related to thyroid cancer [131].

Immunosupression may result in an abnormal immune system which does not effectively recognise and trigger immune responses targeted for cancer cell. This alteration of the immune system is complex and can be mediated by many of the enriched biological processes identified in Table 4.2. Here, high fold enrichment of processes involved with selection and development of T-cells, regulation of humoral immune system processes, immunoglobulin and antigen functionality and those high enrichment of processes related to the ER support the documented relation between suppressed immune system, chronic ER stress and carcinogenesis of secretory tissue [125].

Biological processes relating to T-cells of the immune systems are numerous. Pathways of the immune systems are normally enriched in thyroid gland tissue, because of their involvement in production and regulation of thyroid hormone. Thyroglobulin is a thyroid auto-antigen which provides a matrix in which the synthesis of thyroid hormones take place [136] and variations in the amino acid sequence of this antigen is associated with autoimmune thyroid disease [157]. When thyroid hormone stimulates its receptor in thyroid tissue, the tissue expresses a sodium/iodide symporter (NIS) which imports iodine from the blood stream needed for thyroid hormone synthesis. In some tissues of differentiated thyroid, NIS fails to be correctly placed on the cell surface but localizes in the cytoplasm of the cell [146]. Intracellular retention of NIS have been demonstrated to support cell migration [158]. Because there is significant fold enrichment of immunity related processes reported by the differentially expressed genes, there is indication of abnormal activation and behaviour of genes involved in antigen and immune system pathways in the carcinogenic thyroid tissue. This is supportive of proposed mechanisms by which thyroid cancer may become so invasive [125]. Differential expression analysis and functional enrichment facilitated detailed insights into the molecular mechanisms of thyroid carcinoma.

Immunosupression may result in an abnormal immune system which does not effectively recognise and trigger immune responses targeted for tumor cells. This sub-optimal alteration of the immune system is complex and can be mediated by many of the enriched biological processes identified in Table 4.2. Here, high fold enrichment of processes involved with selection and development of T-cells, regulation of humoral immune system processes, immunoglobulin and antigen functionality and the high enrichment of processes related to

the ER support the documented relation between suppressed immune system, chronic ER stress and carcinogenesis of secretory tissue [125]. Many of the genes found in the network are genes of which altered genetic sequence of expression regulation are associated with thyroid cancer, as presented in Table 4.6. As is evident from the low connectivity of these genes, they do not co-express with many other genes. This suggests that they may have distinct functional properties within the cell and that their own transcription levels are not regulated collectively with several others. These genes are indeed involved in very specific tasks in processes which are very influential of the cell fate. Many of them are regulators of programmed cell death, or apoptosis, regulators of proliferation, cancer cell motility and invasiveness [159].

This disease analysis also aimed to examine the presence of informative content in the network as well as gene modules consisting of genes with significant relation to the disease. The resulting CSD networks differentiated well between the transcriptional interaction patterns in thyroid gland tissue between persons with thyroid cancer and healthy persons. The inferred significant change among genetic associations between the conditions produced several disease-associations, in which both sicknesses of the immune system and cancer were present (Table 4.3). These sicknesses could be traced to network regions dominated by differentiated or specific changes in gene associations. Hence, by contrasting the differences in co-expression, previously uncharacterized information about disease-specific transcriptomic patterns thus became apparent.

The 22 disease-associated DEGs have profound importance for the thyroid cell vitality, see the list in A.1. They roughly cluster into two groups. One major group are cell cycle regulators involvement in programmed cell death, differentiation, transcriptional regulation, and cancer metastasis. An interesting observation that in persons diagnosed with thyroid cancer, their BRAF-gene frequently has been found to harbour genetic alternations and the thyroid tumors of these patients have a tendency of aggressive properties.[18].

The other group consists of genes with major influence in immune system recognition, regulation and conduct of humoral immune responses. Immunosupression may result in an abnormal immune system which does not effectively recognise and trigger immune responses targeted for cancer cell. This alteration of the immune system is complex and can be mediated by many of the enriched biological processes identified in Table 4.2. High fold enrichment of processes involved with selection and development of T-cells, regulation of humoral immune system processes, immunoglobulin and antigen functionality and those of processes related to the ER was observed. This is in support of the documented relation between suppressed immune system, chronic ER stress and carcinogenesis of secretory tissue [145]. These general observations are likely to be representative of the outcomes in the analysis of this thesis. The differential co-expression patterns for the disease-genes were significantly shifted when going from normal transcription patterns to that of thyroid cancer. The ability of the CSD framework to elucidate these differential co-expression categories was a major benefit facilitating biological interpretation. Ultimately, there was strong evidence for their collective aberrant behaviour driving the molecular pathogenesis.

A similar property to immune disease-associated genes is that the cancer-associated cells are also dominated by differential and condition-specific manners of co-expression to other genes. Many of the disease-associated genes are characterized by high average shortest path lengths to other genes. Some of them, e.g. FAS, ADGRV1, DIO1, ITPR1, ITGA3, MATN2, and NCOA4 have very high average shortest paths and are all found in less dense areas of the network making connections between neighborhoods that are far apart. These genes may be influential in transmission of information in the cell; connecting cellular processes together in order to orchestrate a synchronized behaviour. The reversal or sudden presence of co-expression profiles of these genes and others is clear from their predominantly differentiated or specific type links. Strong changes into opposite signed correlation in transcription with other genes and condition-dependency in other strong correlation in transcription is suggestive of an vastly altered regime of cell transcriptional regulation. The change in information flow facilitated by these genes acting abnormally may thus start to act as sources of cancer manifestation in the cell. Changes in transcriptional profile relationships like these are descriptive of dysfunctional behaviour with large deleterious consequences on the metabolic network of the thyroid cells.

A healthy thyroid gland is a highly specialized secretory tissue with very specific processes for synthesis and regulation of thyroid hormone in the human body. Many genes are normally expressed here at high levels for the proper function of the thyroid. To distinguish relationships between DEGs with *malignant characteristics*, investigation of specific genes and modules for association to thyroid cancer was pursued. Biological process enrichment analysis was employed to search for disease-modules. This had to be a network clique with high enrichment for thyroid cancer-associated genes as DEGs who's transcriptional discordance was related to carcinoma. The Louvain-algorithm was employed to partition the network into community-like structures, and the results showed that many of these modules segregated together with regions of common link types, i.e. regions with genes of similar category of association change between the studied conditions (Fig. 4.7). The modules dominated by differentiated and specific type links were investigated further, because associations between the gene-pairs in these either had opposite signs between the compared conditions, or the associations were present in only one conditions exclusively. There is opposite correlation in these genes' transcriptional level between thyroid cancer and normal persons, as indicated by the differentiated (red) links. This reversal is likely to have a tremendous impact on cells, where shift in regulation of genes involved with the ER may lead to inflammation and carcinogenic behaviour in the thyroid tissue.

The identified thyroid-cancer associated DEGs formed a basis for both investigation of molecular pathogenesis of thyroid cancer, and as means of qualitative assessment of network inference results. As explained in Chapter 5.4.4, the purpose of the differential gene co-expression networks generated with the CSD method was to describe correlations of genes, and not highlight the most DEGs per say. It should be expected that certain inferred DEGs in the CSD networks will coincide with individually identified THCA-associated genes. These disease genes were thus especially interesting and were looked for in commu-

nities of the network as well. Several identified network modules contained thyroid-cancer genes. The largest of them, module 5 and module 34 were promising candidates as a disease modules.

Several modules were abnormally enriched for processes of the immune system activation and regulation, inflammation and ER stress, and programmed cell-death (Tables 4.8, 4.9, 4.10, 4.11). These modules are likely manifestations of the carcinogenic phenotype and the genetic association patterns observed among them, virtually only differentiated and specific type links, are supportive of this observation. The genetic interactions in these modules are descriptive of a significant difference of transcription level correlation between then that portray malignant regulation of genes or genetic alterations causing anomalies in the transcriptional regulatory system. Especially interesting is module number 34. Here there is an extremely high enrichment of calcium-ion sequestering, this happens when the endoplasmatic reticulum (ER) is under stress and may lead the mitochondria to induce programmed cell death [160]. This exact process has over a hundred fold enrichment in this module, and all genes involved are interacting with differentiated links. Antigens are also produced and processed in the ER. In addition, pathways related to the immune response, such as antigen processing and presentation through the MCH (major histocompatibility complex) protein, have abnormal high enrichment and support this observation. Moreover, protein folding catabolism and response to hypoxia are symptoms of malfunctioning ER as well. These two in combination are often activated at the same in tumor cell [161]. Lastly, the NIK/NF-kappaB signaling process is involved with regulation of cell survival and inflammation. Activation of this pathway has been reported to mediate pro-inflammatory signals in the cell and antiaptototic pathways [162]. This could hinder programmed cell death in tumour cell and lead to tumor progression.

The CSD method proved to yield biologically meaningful and informative differential gene co-expression networks which readily identified gene pairs for which correlation in transcriptomic levels were significantly different between the two studied conditions. An important aspect of this method is that it's agenda is to investigate the *differential correlation of pairs* of genes, and not the differential co-expression of separate genes. An important limitation to this method is therefore that it may miss some genes that are significantly differentially expressed between conditions because these do not necessarily co-express with other genes to a high enough degree to be included in the inferred network after the importance value threshold is applied. Biologically there are several ways a gene could have relevance to a specific phenotype whilst lack a high transcription correlation pattern with other genes. This could result from transcriptional regulation of high specificity. Transcription factors with high affinity for their target gene sequence will bind and regulate only their targets, which could be only one or a small set of genes. These will thus not correlate so much with the transcriptional levels with other genes because they are separately regulated by an independent regulatory system. Another explanation could be that a DEG could have a very specific role in the cell. Perhaps it is involved in one single biological pathway and if any genetic alteration events should take place in this gene, it would then perform ab-

normally and mediate a diseased phenotype downstream its designated pathway without directly correlating in transcription levels with many other genes. As demonstrated in section 2.1, the union of CSD network analysis of gene-pair transcriptional correlation patterns was a good foundation for the subsequent investigation of important DEGs associated with relevant biological processes or diseased phenotypes. Employing an integrated analysis of the gene expression profile correlation in the thyroid tissue transcriptomic network combined with investigation of interesting genes together resulted in important insights into the pathogenicity of thyroid cancer.

## 6.2 Method development

The second main aim of this thesis was to develop potential methodological improvements that could be applied to the original implementation of the CSD method for differential co-expression network inference. Several aspects of the differential co-expression analysis were studied in detail. This was a great way to gain knowledge in multiple different bioinformatics tools and computer science, as well as the requirements needed to develop and integrate new alternative improvements into an established workflow.

The differences of transcriptomic data sets resulting from different pre-processing were explored. Comparisons of pre-processing effect on transcriptomic data have shown to affect both the connectivity and the compliance of central genes in the inferred networks. The comparison was done on alternative procedures in filtering out low-count genes and batch effect correction. As the results in this section have pointed out, differential gene co-expression network inference is a strong tool when provided data of high quality. It is not however robust to skewed data and technical artifacts, which will have an effect on the inferred networks. When provided a data set of raw count data with potentially high levels of noise and batch effects, an inferred differential co-expression may result in some false gene associations. Given that the raw counts from RNA-seq transcriptomic quantification are very dependent of sequencing depth and gene length this might lead to technical bias in the results.

As demonstrated, lack of pre-processing resulted in a network with systematically higher number of edges and consequently higher average degrees and shorter distances between nodes in the networks. Because these properties may originate from transcriptomic data which is offset toward the high condition-specific expression for some genes, it seems likely that some of the inferred associations in these networks are more likely to be falsely identified. The two other analyses, An. 1 and An. 2 are less likely to suffer from such distortions. The comparison of differential gene co-expression networks on the basis of different quality-control procedures were compared by their qualitative content; the overlap of hubs and ability to infer thyroid cancer-associated genes as DEGs. From the first comparative study the compliance of central network genes were highest among networks based on the strictest quality control. Anyway, zero quality control which loosened the gene co-expression exclu-

sion process produced a larger network with uniquely identified thyroid cancer-associated genes. Comparisons of pre-processing effect on transcriptomic data have shown to affect both the connectivity and the compliance of central genes in the inferred networks. As pointed out, lack of pre-processing resulted in a network with systematic increased number of edges and consequently higher average degrees and shorter distances between nodes in the network based on a set of rational networks.

In addition to the already established CSD workflow based on Spearman's rank correlation coefficient, three additional alternative similarity measures as choices to base this method on were developed. These were all employed to construct differential gene co-expression networks with foundation in all four alternative similarity measures. An important application of the expansion of the standard CSD network with respect to similarity measures, was that the influence each of them had on the resulting differential gene co-expression networks could be evaluated both with graph theoretic tools and biological interpretation.

Firstly, the weighted topological overlap was incorporated into the CSD workflow. Creating a faster and more optimized software was attempted to speed up and expand some algorithms forming the basis of differential gene co-expression analysis. Initial difficulties implementing multi-threading for the computation of weighted topological overlap were most likely due to improper scheduling and large computational overhand. Alternative measures were used to infer networks with CSD on the basis of weighted topological overlap anyway. The network based on weighted topological had interesting structural organization. Because a high ratio of hubs were connected to other hubs, the network had resemblance of assortative tendencies. Most biological networks have dissassortative behaviour, increasing their stability and robustness [163]. Assortative network properties may be masked by structural dissassortativity rooted in their scale-free property [1], and for many biological networks there have been reported dichotomous degree correlation patterns [164]. A differential transcriptome network with these characteristics depict a robust and molecular interplay yet it's well connected structure ensures that information flows faster and more efficiently between different regions. Another interesting structural feature was that the gene-pair associations types segregated into well-defined regions of the network. Homogeneous network neighborhoods indicated that the wTO was able to emphasize and distinguish the various types of associations regimes among gene pairs in the network.

The network based on weighted topological overlap had very interesting hubs. They had substantially higher degrees in this network than the hubs in the standard CSD network. Here, there were 65 genes who's degree was larger than 40 - which is three times as many than in the standard network. Table B.1 shows that the network hubs are *extremely* enriched for important cell pathways already associated to the pathogenicity of autoimmune thyroid disease and thyroid cancer [131, 145]. Many hubs were found in the network components characterized by specific type associations to other genes, indicating the most central players in the transcriptional network are behaving very differently in thyroid glands with cancer compared to healthy tissue.

Some thyorid cancer-associated genes had different connections in the wTO-network than in the conventional CSD network. A good example of this is the gene S100A10. This is found in a small network component consisting of four genes connected by specific type links; S100A10, ALOX5, CATSPER1 and MXRA8. These four genes encode proteins that have various roles in immune system processes: S100A10 protein regulates phosphorylation of ANXA3 which is target for tyrosine-specific kinases, ANXA3 is directly linked to the gene ADGRV1 also identified as a thyroid cancer gene in the table. ALOX5 protein is responsible for biosynthesis of proinflammatory leukotrienes [165], CATSPER1 protein is a calcium channel which is important for regulation of calcium storage in the endoplasmatic reticulum. MXRA8 is a pro-inflammatory protein involved in regulation of cell proliferation and the hedgehog signaling pathway, a pathway suggested to stimulate thyroid cancer cell motility and invasive behaviour [125]. In the normal CSD network this had only one neighbor whereas in the wTO-network the same gene formed an interesting module together with others of high relevance for the studied disease.

All together these results suggests that weighted topological overlap as similarity measure produced a highly integrated network with very interconnected cores consisting of clusters where genes of similar biological functionality form associations. The network segregated well between types of differentiated transcriptomic behaviour. The network hubs represent genes with important roles in organization of cellular processes in the cell. The misregulation observed in the transcriptomic shift during thyroid carcinoma are likely to be explanatory of thyroid cancer pathogenesis.

The second major development part of this thesis was the successful development of the software for mutual information-based network inference with the CSD method. A better implementation of parallelized programming was realized for the estimation of mutual information (MI) as an alternative to the ranked correlation coefficient. Computing differential gene co-expression measures on the basis of MI performed well and generated networks characteristic of complex biological systems, although it's structure was substantially different from the other CSD networks (Fig. 5.5). This strategy nevertheless generated rational networks which provided new insights into the condition under study.

Analysis of enriched processes among the DEG associations in the MI-based CSD network was very interesting, these showed enrichment of yet additional pathways of the immune system and of translational regulation. Enrichment of translational initiation can be characteristic of loss of proper cell cycle regulation, indicative of abnormal correlation among the associated genes [159]. In addition, enrichment of *nonsense-mediated decay* (NMD) can also be a symptom of loss of homeostasis and strong cellular stress levels. Enrichment of this pathway can be rooted in abnormally high transcriptional activity and it may lead to increase in proteins of malignant functionality. The NMD pathway is a mechanism of transcriptional damage-control where aberrant transcripts containing non-sense mutations are targeted and degraded, which in some cases may result in deleterious gain-of-function properties of the transcripts if they are translated into proteins[166]. In the ER proteins are processed after translation, but under strong stress it's performance is obstructed leading

to *unfolded protein response* (UPR). When the ER is chronically stressed, as is typical of autoimmune thyroid disease, NMD is inhibited by UPR [167]. This promotes pathogenesis of autoimmunity and is linked to cancer development and progression [145]. Ultimately, these transcriptome process enrichment on the basis of mutual information are novel in the scope of this thesis, and are supportive of the proposed relation between auto-immune thyroid disease and carcinogenic behaviour [131, 136].

Differential co-expression analysis of biological data is always based on a similarity measure as fundamental strategy to elucidate patterns of coinciding expression of genes across multiple samples. Therefore the chosen similarity measure has an important effect on the resulting co-expression profiles generated. The results from Chapter 5 testify that the similarity measure had a large influence on the outcome of the analysis. The most important *quantitative* assessment of the alternative similarity measures was done by robustness analysis, results for which is presented in Figure 5.8. The synthetic low sample size simulated differential gene co-expression results from applying the CSD workflow based on the four similarity measures CSD, CSD-VAR, wTO, and MI. The overlap coefficient quantified the degree of fidelity in gene associations inferred in sets of decreasing sample size compared to the baseline CSD network inferred from the full data set. This plot shows several important differences between the four similarity measures.

First, the plot shows that wTO performs the best in this robustness comparison. It proved to be the least dependent on sample size across link type. This supports recent observations in [118], where the topological overlap similarly performs well in overlap tests for data sets of decreasing sample size. This effect could be rooted in the approach by which the wTO assimilates information from neighboring nodes for a gene pair into the value of this gene-pair's association. In this way, it takes into account topological features of the nodes in the close vicinity of a pair of correlated genes, which is less likely to be as sensitive to noise as the simple correlation value itself would be. In the plots in Fig.5.8, the overlap coefficient data points for wTO are higher than those from conventional CSD and CSD-VAR. This indicates that wTO applied as similarity measures provides important benefits when studying gene expression data of limited sample size. It also improves the relative fidelity performance of differential gene co-expression networks inferred with the CSD framework.

Second, the figure shows that in this analysis MI is very sensitive to decreasing sample size and is thus most applicable to data sets of adequate size. As clearly seen by the yellow crosses in plots of Fig. 5.8, for all types of gene association relationships, either conserved, differentiated or specific, the MI-based networks has the lowest fidelity of inferred links to the reference network. This could be due to the fact that the variance in mutual information between two random variables follows an asymptotic function dependent of the number of random variables (genes). It is thus recommended to estimate the minimal number of samples needed for accurate estimation of the mutual information based on the number of genes in the data set, and utilize data sets fitting these requirements for differential co-expression network generation [168].

Another explanation of the poor sample-size robustness performance of the MI in this thesis is that in the interest of time, computation was performed with only 20 bootstrapped versions for estimation of variance of the MI. This variance was used for correction of each gene-pair association score, and thus could alter the inferred associations substantially. Hence, observations done on the performance of MI in this thesis are speculative and an in-depth investigation with higher number of bootstrapped MI-scores and multiple data-sets of even larger baseline sample-size is needed. Third, CSD-VAR performs better for lower sample size than conventional CSD. Here, the less stringent inference threshold could explain a higher probability of retaining similar inferred gene associations even in smaller sets, because in general more of any association are included in the network. This could support the application of both methods in the case of expression data set of small sample size, but also in any case if limited time is available. Because the CSD-VAR skips the most time-consuming step of the CSD workflow, these results are promising for differential gene co-expression analysis of large amounts of data or testing purposes.

Fourth, Fig. 5.8 demonstrate that there is a pronounced difference in robustness of inferred gene correlations of the different types. Plot d) shows that especially differentiated type links are less overlapping in networks based on gene expression data of decreasing sample size, whereas conserved associations are the most overlapping. This is likely to be rooted in the difference in how these association type scores are computed.

All together, all these results report a general relation between data set sample size and the overlap of inferred associations compared to a larger reference set. Some general observations were that wTO performs well even for small data sets and the quality of MI is more dependent on sample size. Even though these were quite noticeable trends the validity of the comparison is limited because it is based on one single set of data from each condition and the reference data set did not contain an exceedingly large number of samples either.

The major *qualitative* examination and validation strategy of CSD networks based on alternative similarity measures was facilitated through inspection of their respective ability to identify thyroid cancer-associated genes. This analysis was done for all four alternative similarity measures and the results are summarized in Table 5.8. Implementation of the CSD method without variance-correction resulted in the largest set of thyroid-associated genes identified among the correlating DEGs. This could be due to the looser requirement of including associations in this network inference strategy, because it does not divide the association scores by the variance in correlation. Thus, all values for C-, S-, or D-scores for any pair of genes are higher and the resulting network was larger compared to the conventional CSD method. Even though this could result in a higher ratio of falsely positive correlations between pairs of genes identified as significant, it proved to have a very high ability to include relevant genes for the studied condition. With the significantly reduced computation time needed for CSD-VAR compared to normal CSD, observed high level of ability to identify disease-related genes in the CSD-VAR network is promising. This trade-off suggests that analysis of a transcriptomic data set with both CSD and CSD-VAR could be advantageous depending on the research goal.

The weighted topological overlap as similarity measure for CSD network identified *few* thyroid cancer-associated genes and all of them overlapped with THCA-associated genes identified by networks based on other similarity measures. In this aspect, the wTO did not perform so well. This effect is likely be due to the aforementioned possibility that some THCA-associated genes have limited transcriptional correlation with other genes, and are not typically the most strongly correlated with other genes. Indeed, in networks inferred on the basis of all alternative similarity measures have indications of THCA-associated genes in less densely connected network neighborhoods and the majority are of low degree (See Fig. 4.6 for CSD, Fig. 5.7 for CSD-VAR, Fig. 5.3 for wTO, and Fig. 5.5 for MI). It could also indicate that wTO similarity measure is more likely to miss important genes of lower degree, because it accentuates genes with high-degree by weighting it's links by the summed number of interactions. As seen in Table 4.6, most THCA-associated genes were of low-degree. This means that even genes of low degree can have a huge influence of the cellular system's performance and resulting also the phenotype of the organism.

It should be noted that the MI-based CSD network performed perhaps the best in the qualitative assessment where the ability to include THCA-associated genes among the significant gene correlations was examined. Table A.1 summarizes the four method's ability to identify differentially co-expressed genes in the CSD networks and supports this observation. Basing the CSD framework on this novel implementation of MI as similarity measure, resulted in the largest set of identified THCA-associated genes overall. But as Table A.1 shows, not only did it find the highest number of disease-genes, it also found the highest number of disease-genes not included in those from the other networks. As is given in Table 5.8, the MI resulted in the highest uniquely identified THCA-genes, it found 11 unique genes related to thyroid cancer with significant relationships to transcription levels of other genes. Conversely, the CSD-VAR also identified many THCA-associated genes, but as apparent in the table there was substantial overlap between conventional CSD and CSD-VAR. The MI-network identified over twice as many unique THCA-genes, some of which were not included in any network based on other similarity measures. These are listed in A.3 and are involved with regulation of gene expression, homeostasis and inflammation, antioxidant proteins, and tumor cell proliferation and migration. Many of the disease-associated genes were associating with other genes in the transcriptional network in differentiated or specific manner, indicating that the expression levels of these genes were significantly altered between the studied conditions. This indicates that the MI as similarity measure does capture different forms of associations than the correlation-based methods and expanded the molecular insight into thyroid carcinoma. Conclusively, the results from the qualitative assessment of the similarity measures' effects on the inferred networks demonstrated that both mutual information and the simplified version of CSD did indeed identify unique thyroid cancer-associated genes among significant associations of DEGs. These were not previously included in the networks generated based on the standard CSD method and provided new important insight into the transcriptional traits of pathogenesis of thyroid cancer.

Although the results of this section may seem persuasive, there are some limitations to the

research strategy. Obviously, all conclusions drawn on the quality of the similarity measures are restricted to their performance for the data set investigated in this particular thesis. They were only applied to one empirical data set of TGCA data for thyroid cancer and GTEx data for normal tissue. Investigation of performance of various similarity similarity measures should be done on multiple sets from multiple studied condition in order to draw valid evaluations. The similarity measures have been compared among each other by both qualitative and quantitative means. The similarity measure investigation was thus an attempted comparison on the basis of this single data set and condition under study.

A potential confounding factor in the disease analysis of this thesis became apparent through the analysis of enriched processes associated to the CSD networks. The data set from the The Cancer Genome Atlas contained transcriptomic files measured in thyroid tissue of patients diagnosed with various kinds of thyroid carcinoma. The complete collection of available gene expression data was downloaded from there in order to obtain the largest sample size as possible. As explained in section 3.1.1, this data set consisted of predominantly transcriptomic data from thyroid adenocarcinoma, but there were also several files from patients diagnosed with other types of thyroid cancer. This could be the origin of some observed indications of conflicting observations among results from biological process enrichment analysis. In some cases, such as in section 5.2.1, there were indications of impeded immune system functionality, which was related to published literature linking this to invasive properties of some thyroid cancer sub-types. At the same time, there were also indications of immune system activation in other biological process enrichment results. This was indicative of an immune system perhaps properly recognising and in combat with the tumor cells in the thyroid tissue. Papillary thyroid carcinoma is an innocuous thyroid cancer type, for which these characteristics reflect tumor-induced immune system activation. In addition, certain important differentially co-expressed genes have documented influence on which characteristics the thyroid cancer develops, such as STAT3, FN1 and AMOT. As example, STAT3, is linked to many cancer types and it's expression levels are predictive of tumor aggressiveness or quiescent nature [139]. These DCGs pinpoint promising areas of the CSD network to look for novel prognostic markers, but their respective co-expression correlations to other genes may be subject to ambiguous underlying transcriptomic signals. The mixed nature of pathogenicity of these tumours affected the outcomes in the analyses of biological process enrichment, and is also likely to perplex other aspects of the differential gene co-expression analysis more difficult to detect. Thus, it should be noted that rather composing a thyroid cancer transcription data set of *one single sub-type* of thyroid carcinoma, would improve the reliability and validity of the observed transcriptional changes in the system. A more homogeneous profile of pathogenic perturbation would be more beneficial because it could be more accurately quantified by gene co-expression analysis and investigated. This wold improve the legitimacy of biological interpretations based on the patterns in the CSD network, as well as any potential new important insights or clinical applications.

A second major limitation to this study is the lack of independence between the two main

agendas of this thesis. The first section assumes that there *is* a significant amount of true differentially expressed genes between thyroid cancer and normal thyroid tissue. The elaborations and examinations in the second part is then necessarily based on this same assumption. If this should prove to be an invalid assumption, all the demonstrations in the second part of the thesis will lose ground. The comparisons of alternative similarity measures requires the assumed significant difference between the transcriptomic data. If this was the case, this empirical data set is not fitted for neither differential gene co-expression analysis or for comparison of similarity measure performance. It should be noted that there are several observation of those elaborated in the Results section of this thesis that affirm the assumed differential expression between thyroid cancer and normal thyroid tissue. Conclusively, this is in support of the demonstrated differences among the similarity measures investigated in this thesis.

# Chapter 7

# Conclusion & Outlook

Two sets of gene expression measurement data from thyroid cancer patients and normal controls were compared by differential gene co-expression analysis with the CSD framework. The goal of this part was to perform an in-depth investigation of differentially co-expressed genes between thyroid cancer and normal tissue of the thyroid. The CSD framework successfully produced biologically meaningful representations of comparative gene correlations between the two conditions. Several of the hubs that co-expressed with other genes predominantly in differentiated manner between normal and cancer tissue, had interesting molecular roles and were found between denser network regions, mediating information flow from various corners of the network. The largest hub, MICAL3, and the hub AMOT, are both involved with regulation of cell division, EVC regulates cell differentiation, the long non-coding RNA molecule encoded by FAM111A-DT is likely to regulate the expression level of several genes. The fact that these had reversed associations with their neighbor genes from one condition to the other and the documented essential role they play in central pathways of the cell suggested that these hubs could potentially be novel candidates as prognostic markers of thyroid carcinoma.

Investigations of heterogeneous nodes in respect to association type distribution identified genes marking regions where many genes stopped correlating with normal patterns (conserved links) and started correlating in reversed or condition-specific manner. Investigation of link-type heterogeneous nodes such as CALR, pin-pointed sites of abnormal gene co-expression in genes taking part in cellular processes descriptive of a malfunctioning system. GO enrichment analysis reported significant enriched of biological processes that have previously been documented to be anomalously regulated in thyroid carcinoma. This demonstrated that link distribution of nodes could be employed to identify genes with high relevance to the disease under study. The work presented in this section will hopefully contribute to novel knowledge and point towards novel directions for relevant biological studies of predispositions for thyroid carcinoma.

In the second major part of this thesis the foundation for the CSD framework was investigated, both effects of passing expression data through different pre-processing work-flows

and basing the framework of various similarity measures. The goal was to expand the foundation for the established framework for differential gene co-expression analysis, the similarity measure, and look into what influence this had on the aspects of differential co-expression network quality and potential biological merits gained by any alternative similarity measure.

Because differential gene co-expression analysis requires that RNA-quantification from different experimental should be compared directly, it is recommended to perform sample bias and batch effects correction. Network construction of multiple versions of the same data set with various degrees of filtering and normalization procedures applied, resulted in networks of relatively different character. Comparison of network parameters and content, such as similarity in genes identified as network hubs and overlap of inferred significant gene associations (links), was quantified among the three differently pre-processed sets. This confirmed that pre-processing has pronounced effects on the outcome in the differential gene co-expression network. The inferred CSD networks showed that proper pre-processing, including correction for gene lengths and differences in RNA-seq library composition, as basis for differential gene co-expression analysis with the CSD network will yield results with high-quality suitable for comparative studies. Fortunately, applying these additional pre-processing steps were relatively straightforward, supporting the use of recommended pipelines to address the computational challenges associated with RNA-seq transcriptomic data.

The established CSD method for differential gene co-expression analysis was expanded to potentially elucidate a wider range of genetic associations, beyond those captured by the Spearman correlation coefficient. These alternatives were realized either by applying already existing tools for calculation of similarity measures and adapting them into the tool chain, or by writing scripts for quick computation of novel measures to expand the possible strategies further.

The implementation of the weighted topological overlap (wTO) was motivated by the fact that this similarity measure was developed to produce high-quality similarity measurements between genes, especially applicable for differential gene co-expression analysis. The CSD networks constructed which were based on wTO accentuated relevant genetic correlations by incorporating properties of each gene-pair's connections to their respective neighbors. Adding this topological information as an improvement to the standard correlation to infer gene associations, generated networks which organized the transcriptomic system hierarchically into biological meaningful modular structures. These brought interesting new aspects to the outcomes of the biological process enrichment analysis of the disease under study. Comparisons of fidelity of the wTO-based networks inferred with the CSD method showed that it was the most resilient to simulated low sample-size. Compared to mutual information (MI), Spearman's rank correlation, and the latter without internal variance correction, wTO performed best in respect to inferring similar gene correlations as the reference set. Because it systematically improved the CSD networks robustness performance, applying this transformation of weighted topological overlap measure from the Spearman correlation measures could potentially enhance the quality of these differential co-expression networks

further especially when few experimental measurements from either conditions are available.

As alternative to linear detection methods for inferring genetic correlations, the mutual information (MI) has recently been used to measure similarity in gene co-expression patterns in the pursue of detecting other association patterns than merely those measured by correlation. The MI quantifies the degree of statistical dependence between genes expression vectors based on entropy. In the work of this thesis, an algorithm implementation of this alternative non-linear similarity measure into the established CSD framework was incorporated in a stream-lined way. Implementing it in parallelized manner allowed estimation of MI for gene expression data of large sizes with relatively low execution time. This new workflow facilitated unraveling of novel important components of the complex transcriptomic machinery regulating cellular behaviour.

Assuming that disease-associated genes would coincide as significantly co-expressed with other genes in thyroid tissue as a result of deliberate regulatory mechanisms or physical gene interactions, CSD networks generated on the alternative similarity measures were compared by their relative ability to identify disease-associated genes as significantly co-expressed with other genes. The respective ratios of thyroid cancer-associated genes (THCA-genes) present in CSD networks were identified by Gene Ontology analysis and served as a means of qualitative evaluation, in lack of a validated reference differential co-expression network characterizing the transcriptome of thyroid cancer. Employing MI resulted in uniquely identified THCA-associated genes as significantly co-expressed in the CSD network. This enhanced ability of the MI is likely to be rooted in the theoretical foundation for inferring significant relationships with MI, which is based on a completely separate estimation method than the correlation-based ones. On the other hand, CSD networks based on this more far-reaching similarity measure were more dependent of sample size. Based on the results of this thesis, the applicability of MI as similarity measure required that the gene expression data has an adequate number of samples and that the MI-based implementation of the CSD framework should be run with more than 20 bootstrap samples of the gene expression data.

Several suggestions to future work apply for the first aim of this thesis, the disease investigation part. Investigations performed on the CSD network on node homogeneity resulted in identification of several interesting *heterogeneous* genes, in respect to their link type distributions. These marked the transitions from normal to diseased patterns in the transcriptomic network and are promising candidates of disease genes. Experimental investigations of the biological functions and effect of erroneous regulation in these heterogeneous genes could potentially reveal strong relevance to the disease under study. The same applies to modular structures and hub genes in these networks. Especially network hubs that were dominated by specific links or differentiated links were extremely enriched for cellular processes descriptive of malignant regulatory behaviour. The hubs' expression levels significantly correlated with *numerous* other genes, suggesting an aberrant regulatory role in the cells which would be very interesting to investigate further.

As for the method development part, application of the weighted topological overlap mea-

sure improved the fidelity of the CSD networks and provide a useful improvement, especially when studying gene expression data sets of limited sample size. The developed software for application of the non-linear similarity measure mutual information utilized of a different strategy of inferring interesting co-expression relationships, allowing a wider range of knowledge to be extracted from the CSD network. It thus provides a method for expanded investigation of cellular mechanisms underlying specific phenotypes and diseases. Because it proved to be sensitive to sample size, it's applicability is dependent on estimating a required minimal sample size given the number of genes and basing the differential co-expression networks on data sets of adequate size.

Further research is needed in order to fully characterize the complex interplay of events driving carcinogenic development of thyroid cells into cancer cells. Analysis of the transcriptional associations which correlated in condition-specific and differentiated manner between the normal and cancer tissue could be a next step towards deciphering which components are mediating the abnormal regulatory profiles altering the cellular behaviour. Biological validations of the inferred associations present in the CSD network and the nature of the correlation types would be interesting to study as well. Some of the detected modular structures in the network comprised defined neighborhoods with integrated genetic expression associations that are likely to represent disease modules. It would be interesting to study these further with both systems biology approaches and experimental validations to examine how the expression associations of these genes manifest regulatory mechanisms in the cell leading to disease pathogenicity. This knowledge could elucidate how these associations could potentially be restored back to their normal states, so that the complex interplay of gene associations could be made more resilient to aberrant behaviour of certain components and resolve disease-associated genetic interplay ultimately curing the disease.

# Bibliography


[1] A.-L. Barabási and M. Pósfai, *Network science* (Cambridge University Press, Cambridge, 2016), ISBN 9781107076266 1107076269.

[2] E. Voit, *A First Course in Systems Biology* (Garland Science, 2017), ISBN 9780815345688.

[3] A. Voigt, K. Nowick, and E. Almaas, PLOS Computational Biology **13**, e1005739 (2017), ISSN 1553-7358.

[4] V. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, Journal of Statistical Mechanics: Theory and Experiment **2008** (2008), ISSN 1742-5468, copyright 2008 Elsevier B.V., All rights reserved.

[5] K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang, Nature genetics **36**, 431 (2004), ISSN 1061-4036.

[6] C. Adami, BioEssays **24**, 1085 (2002), ISSN 1521-1878.

[7] A. Trewavas, The Plant cell **18**, 2420 (2006), ISSN 1040-4651.

[8] H. V. Westerhoff and B. O. Palsson, Nature Biotechnology **22**, 1249 (2004).

[9] A. M. Arias and P. Hayward, Nature Reviews Genetics **7**, 34 (2006), ISSN 1471-0064.

[10] . G. P. Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, et al., Nature **526**, 68 (2015), ISSN 1476-4687.

[11] S. van Dam, U. Võsa, A. van der Graaf, L. Franke, and J. P. de Magalhães, Briefings in bioinformatics **19**, 575 (2018), ISSN 1477-4054.

[12] K. Lage, N. T. Hansen, E. O. Karlberg, A. C. Eklund, F. S. Roque, P. K. Donahoe, Z. Szallasi, T. S. Jensen, and S. Brunak, Proceedings of the National Academy of Sciences of the United States of America **105**, 20870 (2008), ISSN 1091-6490.

[13] L. Hood, J. R. Heath, M. E. Phelps, and B. Lin, Science **306**, 640 (2004), ISSN 0036-8075, 1095-9203.

[14] N. Howlader, `http://seer.cancer.gov/csr/1975_2008/` (2011).

[15] J. J. Wiltshire, T. M. Drake, L. Uttley, and S. P. Balasubramanian, Thyroid **26**, 1541 (2016), pMID: 27571228.

[16] M. Xing, Nature reviews. Cancer **13**, 184 (2013), ISSN 1474-1768.

[17] Z. W. Baloch, V. A. LiVolsi, S. L. Asa, J. Rosai, M. J. Merino, G. Randolph, P. Vielh, R. M. DeMay, M. K. Sidawy, and W. J. Frable, Diagnostic cytopathology **36**, 425 (2008).

[18] V. A. LiVolsi, Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc **24 Suppl 2**, S1 (2011), ISSN 1530-0285.

[19] R. E. Gardner, R. M. Tuttle, K. D. Burman, S. Haddady, C. Truman, Y. H. Sparling, L. Wartofsky, R. B. Sessions, and M. D. Ringel, Archives of Otolaryngology–Head & Neck Surgery **126**, 309 (2000).

[20] H. J. Kim, J.-Y. Sung, Y. L. Oh, J. H. Kim, Y.-I. Son, Y.-K. Min, S. W. Kim, and J. H. Chung, Head & neck **36**, 1695 (2014), ISSN 1097-0347.

[21] A. T. A. A. G. T. on Thyroid Nodules, D. T. Cancer, D. S. Cooper, G. M. Doherty, B. R. Haugen, B. R. Hauger, R. T. Kloos, S. L. Lee, S. J. Mandel, E. L. Mazzaferri, et al., Thyroid : official journal of the American Thyroid Association **19**, 1167 (2009), ISSN 1557-9077.

[22] M. E. Cabanillas, D. G. McFadden, and C. Durante, Lancet (London, England) **388**, 2783 (2016), ISSN 1474-547X.

[23] E. Albi, S. Cataldi, A. Lazzarini, M. Codini, T. Beccari, F. S. Ambesi-Impiombato, and F. Curcio, International journal of molecular sciences **18** (2017), ISSN 1422-0067.

[24] L. Manzella, S. Stella, M. Pennisi, E. Tirrò, M. Massimino, C. Romano, A. Puma, M. Tavarelli, and P. Vigneri, International journal of molecular sciences **18**, 1325 (2017).

[25] D. Grimm, International journal of molecular sciences **18** (2017), ISSN 1422-0067.

[26] R.-H. Song, Q. Wang, Q.-M. Yao, X.-Q. Shao, L. Li, W. Wang, X.-F. An, Q. Li, and J.-A. Zhang, International journal of molecular sciences **17** (2016), ISSN 1422-0067.

[27] M. Eszlinger, K. Krohn, and R. Paschke, The Journal of Clinical Endocrinology & Metabolism **86**, 4834 (2001).

[28] Y. Huang, M. Prasad, W. J. Lemon, H. Hampel, F. A. Wright, K. Kornacker, V. LiVolsi, W. Frankel, R. T. Kloos, C. Eng, et al., Proceedings of the National Academy of Sciences **98**, 15044 (2001).

[29] F. H. CRICK, Symposia of the Society for Experimental Biology **12**, 138 (1958), ISSN 0081-1386.

[30] F. JACOB and J. MONOD, Journal of molecular biology **3**, 318 (1961), ISSN 0022-2836.

[31] S. M. Yoo, J. H. Choi, S. Y. Lee, and N. C. Yoo, Journal of microbiology and biotechnology **19**, 635 (2009), ISSN 1017-7825.

[32] T. I. Lee and R. A. Young, Cell **152**, 1237 (2013), ISSN 1097-4172.

[33] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, Genome research **18**, 1509 (2008), ISSN 1088-9051.

[34] K. R. Kukurba and S. B. Montgomery, Cold Spring Harbor protocols **2015**, 951 (2015), ISSN 1559-6095.

[35] G. K. Marinov, Briefings in functional genomics **16**, 326 (2017), ISSN 2041-2657.

[36] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, Nature methods **5**, 621 (2008), ISSN 1548-7105.

[37] M. A. Garcia-Blanco, A. P. Baraniak, and E. L. Lasda, Nature biotechnology **22**, 535 (2004), ISSN 1087-0156.

[38] G. Jin, J. Sun, S. D. Isaacs, K. E. Wiley, S.-T. Kim, L. W. Chu, Z. Zhang, H. Zhao, S. L. Zheng, W. B. Isaacs, et al., Carcinogenesis **32**, 1655 (2011), ISSN 1460-2180.

[39] W. Ge, X. Ma, X. Li, Y. Wang, C. Li, H. Meng, X. Liu, Z. Yu, S. You, and L. Qiu, Leukemia research **33**, 948 (2009), ISSN 1873-5835.

[40] R. S. Sekhon, R. Briskine, C. N. Hirsch, C. L. Myers, N. M. Springer, C. R. Buell, N. de Leon, and S. M. Kaeppler, PloS one **8**, e61005 (2013), ISSN 1932-6203.

[41] H. Richard, M. H. Schulz, M. Sultan, A. Nürnberger, S. Schrinner, D. Balzereit, E. Dagand, A. Rasche, H. Lehrach, M. Vingron, et al., Nucleic acids research **38**, e112 (2010), ISSN 1362-4962.

[42] S. Ballouz, W. Verleyen, and J. Gillis, Bioinformatics (Oxford, England) **31**, 2123 (2015), ISSN 1367-4811.

[43] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, et al., Genome biology **17**, 13 (2016), ISSN 1474-760X.

[44] S. W. Wingett and S. Andrews, F1000Research **7**, 1338 (2018), ISSN 2046-1402.

[45] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, Bioinformatics (Oxford, England) **29**, 15 (2013), ISSN 1367-4811.

[46] S. Anders, P. T. Pyl, and W. Huber, Bioinformatics (Oxford, England) **31**, 166 (2015), ISSN 1367-4811.

[47] M.-A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, et al., Briefings in bioinformatics **14**, 671 (2013), ISSN 1477-4054.

[48] P. P. Labaj, G. G. Leparc, B. E. Linggi, L. M. Markillie, H. S. Wiley, and D. P. Kreil, Bioinformatics **27**, i383 (2011).

[49] S. Sun, M. Hood, L. Scott, Q. Peng, S. Mukherjee, J. Tung, and X. Zhou, Nucleic acids research **45**, e106 (2017), ISSN 1362-4962.

[50] M. D. Robinson and A. Oshlack, Genome Biology **11**, R25 (2010), ISSN 1474-760X.

[51] A. Rau and C. Maugis-Rabusseau, Briefings in Bioinformatics **19**, 425 (2017), ISSN 1477-4054.

[52] O. Wolkenhauer, C. Auffray, O. Brass, J. Clairambault, A. Deutsch, D. Drasdo, F. Gervasio, L. Preziosi, P. Maini, A. Marciniak-Czochra, et al., Genome medicine **6**, 21 (2014), ISSN 1756-994X.

[53] G. A. Pavlopoulos, M. Secrier, C. N. Moschopoulos, T. G. Soldatos, S. Kossida, J. Aerts, R. Schneider, and P. G. Bagos, BioData mining **4**, 10 (2011), ISSN 1756-0381.

[54] L. C. Freeman, Journal of Sociometry **40**, 35 (1977).

[55] M. J. Newman, Social Networks **27**, 39 (2005), ISSN 0378-8733.

[56] D. Iacobucci, R. McBride, and D. L. Popovich, Journal of Social Structure **18** (2017), ISSN 1529-1227.

[57] J. Golbeck, in *Analyzing the Social Web*, edited by J. Golbeck (Morgan Kaufmann, Boston, 2013), pp. 25 – 44, ISBN 978-0-12-405531-5.

[58] W. D. Schank, T, Journal of Graph Algorithms and Application **9**, 265 (2005).

[59] D. J. Watts and S. H. Strogatz, Nature **393**, 440 (1998), ISSN 1476-4687.

[60] M. E. J. Newman, D. J. Watts, and S. H. Strogatz, Proceedings of the National Academy of Sciences of the United States of America **99 Suppl 1**, 2566 (2002), ISSN 0027-8424.

[61] C. O. Daub, R. Steuer, J. Selbig, and S. Kloska, BMC bioinformatics **5**, 118 (2004), ISSN 1471-2105.

[62] I. Priness, O. Maimon, and I. Ben-Gal, BMC bioinformatics **8**, 111 (2007), ISSN 1471-2105.

[63] P. D'haeseleer, S. Liang, and R. Somogyi, Bioinformatics (Oxford, England) **16**, 707 (2000), ISSN 1367-4803.

[64] A. A. Goshtasby, *Similarity and Dissimilarity Measures* (Springer London, London, 2012), pp. 7–66, ISBN 978-1-4471-2458-0, URL `https://doi.org/10.1007/978-1-4471-2458-0_2`.

[65] L. Song, P. Langfelder, and S. Horvath, BMC Bioinformatics **13**, 328 (2012), ISSN 1471-2105.

[66] R. O. Ness, K. Sachs, and O. Vitek, Journal of Proteome Research **15**, 683 (2016), pMID: 26731284.

[67] F. Markowetz and R. Spang, BMC Bioinformatics **8**, S5 (2007), ISSN 1471-2105.

[68] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al., Nature genetics **25**, 25 (2000), ISSN 1061-4036.

[69] C. Spearman, The American journal of psychology **100**, 441 (1987), ISSN 0002-9556.

[70] R. W. Conners and C. A. Harlow, IEEE transactions on pattern analysis and machine intelligence **2**, 204 (1980), ISSN 0162-8828.

[71] A. J. Butte and I. S. Kohane, Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing pp. 418–429 (2000), ISSN 2335-6928.

[72] C. O. Daub, J. Kurths, J. Selbig, J. Weise, and R. Steuer, Bioinformatics **18**, S231 (2002), ISSN 1367-4803.

[73] W. Li, Journal of Statistical Physics **60**, 823 (1990), ISSN 1572-9613.

[74] Y.-I. Moon, B. Rajagopalan, and U. Lall, Phys. Rev. E **52**, 2318 (1995).

[75] A. Kraskov, H. Stögbauer, and P. Grassberger, Phys. Rev. E **69**, 066138 (2004).

[76] G. S. Michaels, D. B. Carr, M. Askenazi, S. Fuhrman, X. Wen, and R. Somogyi, Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing pp. 42–53 (1998), ISSN 2335-6928.

[77] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, Bioinformatics (Oxford, England) **18 Suppl 2**, S231 (2002), ISSN 1367-4803.

[78] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, Science **297**, 1551 (2002), ISSN 0036-8075.

[79] G. S. Davidson, B. Wylie, and K. Boyack (2001), pp. 23–30.

[80] S. L. Carter, C. M. Brechbühler, M. Griffin, and A. T. Bond, Bioinformatics (Oxford, England) **20**, 2242 (2004), ISSN 1367-4803.

[81] S. Horvath, *Weighted Network Analysis: Applications in Genomics and Systems Biology*, SpringerLink : Bücher (Springer New York, 2011), ISBN 9781441988195.

[82] M. K. Vijaymeena and K. Kavitha, Machine Learning and Applications: An international Journal (MLAIJ) **3**, 4 (2016).

[83] L. D. Thomas, D. Vyshenska, N. Shulzhenko, A. Yambartsev, and A. Morgun, F1000Research **5**, 2740 (2016), ISSN 2046-1402.

[84] E. K. S. Joseph Loscalzo, Albert-László Barabási, *Network Medicine: Complex Systems in Human Disease and Therapeutics* (Harvard University Press, Cambridge, 2017).

[85] A. Oshlack, M. D. Robinson, and M. D. Young, Genome biology **11**, 220 (2010), ISSN 1474-760X.

[86] E. A. R. Serin, H. Nijveen, H. W. M. Hilhorst, and W. Ligterink, Frontiers in plant science **7**, 444 (2016), ISSN 1664-462X.

[87] A.-L. Barabási and Z. N. Oltvai, Nature Reviews Genetics **5**, 101 (2004), ISSN 1471-0064.

[88] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, Nature reviews. Genetics **12**, 56 (2011), ISSN 1471-0064.

[89] A. A. Hagberg, D. A. Schult, and P. J. Swart, in *Proceedings of the 7th Python in Science Conference*, edited by G. Varoquaux, T. Vaught, and J. Millman (Pasadena, CA USA, 2008), pp. 11 – 15.

[90] Y. Yang, L. Han, Y. Yuan, J. Li, N. Hei, and H. Liang, Nature communications **5**, 3231 (2014).

[91] C. S. Greene, A. Krishnan, A. K. Wong, E. Ricciotti, R. A. Zelaya, D. S. Himmelstein, R. Zhang, B. M. Hartmann, E. Zaslavsky, S. C. Sealfon, et al., Nature Genetics **47**, 569 (2015).

[92] A. Sharma, J. Menche, C. C. Huang, T. Ort, X. Zhou, M. Kitsak, N. Sahni, D. Thibault, L. Voung, F. Guo, et al., Human molecular genetics **24**, 3005 (2015), ISSN 1460-2083.

[93] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, Proceedings of the National Academy of Sciences of the United States of America **104**, 8685 (2007), ISSN 0027-8424.

[94] S. Ferrari, E. Di Iorio, V. Barbaro, D. Ponzin, F. S. Sorrentino, and F. Parmeggiani, Current genomics **12**, 238 (2011), ISSN 1875-5488.

[95] Z. Tu, C. Argmann, K. K. Wong, L. J. Mitnaul, S. Edwards, I. C. Sach, J. Zhu, and E. E. Schadt, Genome research **19**, 1057 (2009), ISSN 1088-9051.

[96] *Hypothesis Testing* (Springer New York, New York, NY, 2008), pp. 250–252, ISBN 978-0-387-32833-1.

[97] S.-Y. Chen, Z. Feng, and X. Yi, Journal of thoracic disease **9**, 1725 (2017), ISSN 2072-1439.

[98] Y. Benjamini and Y. Hochberg (1995), pp. 289–300.

[99] J. Simes, Biometrika **73**, 751 (1986).

[100] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, et al., Nature Genetics **45**, 580 (2013), ISSN 1546-1718.

[101] D. S. DeLuca, J. Z. Levin, A. Sivachenko, T. Fennell, M.-D. Nazaire, C. Williams, M. Reich, W. Winckler, and G. Getz, Bioinformatics (Oxford, England) **28**, 1530 (2012), ISSN 1367-4811.

[102] B. Li and C. N. Dewey, BMC Bioinformatics **12**, 323 (2011), ISSN 1471-2105.

[103] C. G. A. R. Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, Nature genetics **45**, 1113 (2013), ISSN 1546-1718.

[104] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, et al., Genome research **22**, 1760 (2012), ISSN 1549-5469.

[105] Y. Chen, A. Lun, and G. Smyth, F1000Research **5** (2016).

[106] y. . . p. . G. j. . G. h. . `https://github.com/andre-voigt/CSD`. c. . c. André Voigt, title = CSD.

[107] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, Genome research **13**, 2498 (2003), ISSN 1088-9051.

[108] M. Krzywinski and N. Altman, *Points of significance: visualizing samples with box plots* (2014).

[109] D. W. Huang, B. T. Sherman, and R. A. Lempicki, Nucleic acids research **37**, 1 (2009), ISSN 1362-4962.

[110] P. D. Thomas, M. J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, and A. Narechania, Genome research **13**, 2129 (2003), ISSN 1088-9051.

[111] J. Pontius, L. Wagner, and G. Schuler, Technical Report, National Center for Biotechnology Information (2004).

[112] GTExThe GTEx Portal, *Documentation* (2014), data retrieved from GTEx Documentation Single-Tissue eQTL Analysis.

[113] GeneALaCartGeneCards Bacth Queries, *Genecards batch queries* (2019), data retrieved from GeneALaCart.

[114] Y. Assenov, F. Ramírez, S.-E. Schelhorn, T. Lengauer, and M. Albrecht, Bioinformatics (Oxford, England) **24**, 282 (2008), ISSN 1367-4811.

[115] V. Kumar, A. Grama, A. Gupta, and G. Karypis, *Introduction to parallel computing. Design and analysis of algorithms*, vol. 2 (1994).

[116] D. M. Gysi, A. Voigt, T. d. M. Fragoso, E. Almaas, and K. Nowick, BMC Bioinformatics **19**, 392 (2018), ISSN 1471-2105.

[117] B. Zhang and S. Horvath, Statistical applications in genetics and molecular biology **4**, Article17 (2005), ISSN 1544-6115.

[118] A. Voigt and E. Almaas, BMC Bioinformatics **20**, 58 (2019), ISSN 1471-2105.

[119] A. Voigt, Ph.D. thesis, Norwegian University of Science and Technology, Faculty of Natural Sciences, Department of Biotechnology and Food Science (2018), iSBN 978-82-326-3422-4, Doctoral Thesis at NTNU, 2018:316.

[120] C. E. Shannon, Bell system technical journal **27**, 379 (1948).

[121] L. Paninski, Neural Computation **15**, 1191 (2003).

[122] J. Ish-Horowicz and J. Reid, bioRxiv (2017).

[123] G. Sales and C. Romualdi, *parmigene: Parallel Mutual Information estimation for Gene Network reconstruction.* (2012), r package version 1.0.2.

[124] S. Wuchty, Z. N. Oltvai, and A.-L. Barabási, Nature Genetics **35**, 176 (2003), ISSN 1546-1718.

[125] A. J. Williamson, M. E. Doscas, J. Ye, K. B. Heiden, M. Xing, Y. Li, R. A. Prinz, and X. Xu, Oncotarget **7**, 10472 (2016), ISSN 1949-2553.

[126] B. Troyanovsky, T. Levchenko, G. Månsson, O. Matvijenko, and L. Holmgren, The Journal of cell biology **152**, 1247 (2001), ISSN 0021-9525.

[127] S. Unger, M. W. Górna, A. Le Béchec, S. Do Vale-Pereira, M. F. Bedeschi, S. Geiberger, G. Grigelioniene, E. Horemuzova, F. Lalatta, E. Lausch, et al., American journal of human genetics **92**, 990 (2013), ISSN 1537-6605.

[128] C. Lepoivre, M. Belhocine, A. Bergon, A. Griffon, M. Yammine, L. Vanhille, J. Zacarias-Cabeza, M.-A. Garibal, F. Koch, M. A. Maqbool, et al., BMC Genomics **14**, 914 (2013), ISSN 1471-2164.

[129] A. Fatica and I. Bozzoni, Nature Reviews Genetics **15**, 7 (2014).

[130] B. D. Adams, C. Parsons, L. Walker, W. C. Zhang, and F. J. Slack, The Journal of clinical investigation **127**, 761 (2017).

[131] L. Zhang, X. Cheng, S. Xu, J. Bao, and H. Yu, Medicine **97**, e11095 (2018), ISSN 1536-5964.

[132] A. Tanay, Genome research **16**, 962 (2006), ISSN 1088-9051.

[133] I. Sela and D. B. Lukatsky, Biophysical journal **101**, 160 (2011), ISSN 1542-0086.

[134] L. Klein, B. Kyewski, P. M. Allen, and K. A. Hogquist, Nature reviews. Immunology **14**, 377 (2014), ISSN 1474-1741.

[135] Y. Nyathi, B. M. Wilkinson, and M. R. Pool, Biochimica et Biophysica Acta (BBA) - Molecular Cell Research **1833**, 2392 (2013), ISSN 0167-4889, functional and structural diversity of the endoplasmic reticulum.

[136] S. M. McLachlan and B. Rapoport, Endocrine reviews **35**, 59 (2014), ISSN 1945-7189.

[137] Y. Caballero, E. M. López-Tomassetti, J. Favre, J. R. Santana, J. J. Cabrera, and J. R. Hernández, Thyroid research **8**, 12 (2015), ISSN 1756-6614.

[138] M. Sponziello, F. Rosignolo, M. Celano, V. Maggisano, V. Pecce, R. F. De Rose, G. E. Lombardo, C. Durante, S. Filetti, G. Damante, et al., Molecular and cellular endocrinology **431**, 123 (2016), ISSN 1872-8057.

[139] N. Sosonkina, D. Starenki, and J.-I. Park, Cancers **6**, 526 (2014), ISSN 2072-6694.

[140] M. Ito, S. Nakagawa, K. Mizuguchi, and T. Okumura, in *Current Approaches in Applied Artificial Intelligence*, edited by M. Ali, Y. S. Kwon, C.-H. Lee, J. Kim, and Y. Kim (Springer International Publishing, Cham, 2015), pp. 120–130, ISBN 978-3-319-19066-2.

[141] P. M. Hogarth and G. A. Pietersz, Nature Reviews Drug Discovery **11**, 311 (2012).

[142] C. Zhao, Y. Gao, N. Yu, T. Li, Y. Zhang, H. Zhang, G. Lu, Y. Gao, and X. Guo, Molecular and cellular endocrinology **477**, 103 (2018), ISSN 1872-8057.

[143] S. D. Lucas, A. Karlsson-Parra, B. Nilsson, L. Grimelius, G. Åkerström, J. Rastad, and C. Juhlin, Human Pathology **27**, 1329 (1996), ISSN 0046-8177.

[144] N. Tsuchida, M.-A. Ikeda, Y. Ishino, M. Grieco, and G. Vecchio, International journal of oncology **50**, 2043 (2017), ISSN 1791-2423.

[145] S. Rahman, A. Archana, A. T. Jan, D. Dutta, A. Shankar, J. Kim, and R. Minakshi, Frontiers in immunology **10**, 344 (2019), ISSN 1664-3224.

[146] T. Kogai and G. A. Brent, Pharmacology & therapeutics **135**, 355 (2012), ISSN 1879-016X.

[147] M. J. Holechek, Nephrology nursing journal : journal of the American Nephrology Nurses' Association **30**, 285 (2003), ISSN 1526-744X.

[148] K. Stamatopoulos, C. Belessi, A. Hadzidimitriou, T. Smilevska, E. Kalagiakou, K. Hatzi, N. Stavroyianni, A. Athanasiadou, A. Tsompanakou, T. Papadaki, et al., Blood **106**, 3575 (2005), ISSN 0006-4971.

[149] D. B. Bloch, S. M. de la Monte, P. Guigaouri, A. Filippov, and K. D. Bloch, The Journal of biological chemistry **271**, 29198 (1996), ISSN 0021-9258.

[150] N. V. Morgan, S. Goddard, T. S. Cardno, D. McDonald, F. Rahman, D. Barge, A. Ciupek, A. Straatman-Iwanowska, S. Pasha, M. Guckian, et al., The Journal of clinical investigation **121**, 695 (2011), ISSN 1558-8238.

[151] M. Yano, Y. Koumoto, Y. Kanesaki, X. Wu, and H. Kido, Biomacromolecules **5**, 1465 (2004), ISSN 1525-7797.

[152] C. Dahlgren and A. Karlsson, Journal of immunological methods **232**, 3 (1999), ISSN 0022-1759.

[153] M. Yunta and P. A. Lazo, Oncogene **22**, 1219 (2003), ISSN 0950-9232.

[154] Z. Gu, C. Lin, J. Hu, J. Xia, S. Wei, and D. Gao, Biological & pharmaceutical bulletin **42**, 573 (2019), ISSN 1347-5215.

[155] X. Lou, B. Kang, J. Zhang, C. Hao, X. Tian, W. Li, N. Xu, Y. Lu, and S. Liu, Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease **1842**, 1423 (2014), ISSN 0925-4439.

[156] D. E. Goldgar, D. F. Easton, L. A. Cannon-Albright, and M. H. Skolnick, Journal of the National Cancer Institute **86**, 1600 (1994), ISSN 0027-8874.

[157] Y. Ban, D. A. Greenberg, E. Concepcion, L. Skrabanek, R. Villanueva, and Y. Tomer, Proceedings of the National Academy of Sciences of the United States of America **100**, 15119 (2003), ISSN 0027-8424.

[158] C. Lacoste, J. Hervé, M. Bou Nader, A. Dos Santos, N. Moniaux, Y. Valogne, R. Montjean, O. Dorseuil, D. Samuel, D. Cassio, et al., Cancer research **72**, 5505 (2012), ISSN 1538-7445.

[159] P. Kotsantis, L. M. Silva, S. Irmscher, R. M. Jones, L. Folkes, N. Gromak, and E. Petermann, Nature Communications **7**, 13087 (2016).

[160] A. Deniaud, O. Sharaf el dein, E. Maillier, D. Poncet, G. Kroemer, C. Lemaire, and C. Brenner, Oncogene **27**, 285 (2008), ISSN 1476-5594.

[161] E. R. Pereira, K. Frudd, W. Awad, and L. M. Hendershot, The Journal of biological chemistry **289**, 3352 (2014), ISSN 1083-351X.

[162] A. Loverre, P. Ditonno, A. Crovace, L. Gesualdo, E. Ranieri, P. Pontrelli, G. Stallone, B. Infante, A. Schena, S. Di Paolo, et al., Journal of the American Society of Nephrology : JASN **15**, 2675 (2004), ISSN 1046-6673.

[163] M. E. J. Newman, Phys. Rev. E **67**, 026126 (2003), URL `https://link.aps.org/doi/10.1103/PhysRevE.67.026126`.

[164] D. Hao and C. Li, PLOS ONE **6**, 1 (2011), URL `https://doi.org/10.1371/journal.pone.0028322`.

[165] N. C. Gilbert, S. G. Bartlett, M. T. Waight, D. B. Neau, W. E. Boeglin, A. R. Brash, and M. E. Newcomer, Science (New York, N.Y.) **331**, 217 (2011), ISSN 1095-9203.

[166] Y.-F. Chang, J. S. Imam, and M. F. Wilkinson, Annual review of biochemistry **76**, 51 (2007), ISSN 0066-4154.

[167] A. E. Goetz and M. Wilkinson, Cellular and molecular life sciences : CMLS **74**, 3509 (2017), ISSN 1420-9071.

[168] M. A. Gil, Applied Mathematics and Computation **30**, 125 (1989), ISSN 0096-3003.

[169] Y. Wang, Y. Shen, S. Wang, Q. Shen, and X. Zhou, Cancer letters **415**, 117 (2018), ISSN 1872-7980.

[170] A. V. Rozhkova, M. V. Zinovyeva, A. V. Sass, I. B. Zborovskaya, S. A. Limborska, and L. V. Dergunova, Molekuliarnaia biologiia **48**, 395 (2014), ISSN 0026-8984.

[171] R. Bellelli, G. Federico, A. Matte', D. Colecchia, A. Iolascon, M. Chiariello, M. Santoro, L. De Franceschi, and F. Carlomagno, Cell reports **14**, 411 (2016), ISSN 2211-1247.

[172] X. Ma, Z. He, L. Li, D. Yang, and G. Liu, Oncotarget **8**, 77761 (2017), ISSN 1949-2553.

[173] S. Washizuka, K. Iwamoto, C. Kakiuchi, M. Bundo, and T. Kato, Neuroscience research **63**, 199 (2009), ISSN 0168-0102.

[174] L. Fagerberg, B. M. Hallström, P. Oksvold, C. Kampf, D. Djureinovic, J. Odeberg, M. Habuka, S. Tahmasebpoor, A. Danielsson, K. Edlund, et al., Molecular & cellular proteomics : MCP **13**, 397 (2014), ISSN 1535-9484.

[175] E. D. Accordi, P. Xekouki, B. Azevedo, R. B. de Alexandre, C. Frasson, S. M. Gantzel, G. Z. Papadakis, A. Angelousi, C. A. Stratakis, V. S. Sotomaior, et al., European thyroid journal **5**, 94 (2016), ISSN 2235-0640.

[176] G. Kallifatidis, D. K. Smith, D. S. Morera, J. Gao, M. J. Hennig, J. J. Hoy, R. F. Pearce, I. R. Dabke, J. Li, A. S. Merseburger, et al., Molecular cancer therapeutics **18**, 801 (2019), ISSN 1538-8514.

# Appendix A

# Disease genes

List of thyroid cancer-associated genes (THCA-genes) identified as differentially expressed genes (DEGs) in the CSD networks generated in the development part of this thesis. The source of information about genetic functions is the UniGene Database [111], otherwise citations are provided. The relatedness of these genes to thyroid cancer is documented by the Gene Association Database [5]. The lists provides information about each gene's biological function and may provide information about which genes it is connected to along with the corresponding link types. The links represent associations between DEGs inferred by the CSD framework, which are either conserved, specific or differential between the studied conditions (thyroid cancer vs. normal).

The first list is based on the THCA-genes in Table 4.6 and support the results in Chapter 4. The two following lists are based on results from Chapter 5. A.2 lists relevant information for THCA-genes for the network based on weighted topological overlap as similarity measure from Table 5.4. A.3 lists information about THCA-genes from the mutual information-based CSD network. Each lists of these latter ones will exclusively list the THCA-genes uniquely identified by their respective methods if not already identified by a network based on Spearman correlation as similarity measure. Lastly, Table A.1 contains information about which THCA-associated genes were found by which alternative CSD methods from Chapter 5.

## A.1 Thyroid cancer associated genes in CSD network

- TPO encodes thyroid peroxidase, a membrane bound protein, which is needed for proper production of thyroid hormones [137].
- ITGB2 encodes information for one sub-unit of the $\beta 2$. Integrin proteins are involved in regulation of cellular adhesion and signal transmission pathways controlling cell growth and proliferation. Also involved in immune system as important cell-membrane constituents of leukocytes for recognition of signs of inflammation.

- HLA-DRB1 encodes a histocompatibility antigen, where it plays a central role in humman immunity processes by presenting peptides of extracellular proteins.
- IRS1 encodes an insulin receptor and has normally high expression in the thyroid.
- FN1 encodes fibronectin-1 protein needed for cellular expansion, migration and differentiation.
- The gene CCL5 encodes a secreted protein involved in immunoregulatory and inflammatory processes.
- STAT3 is connected to network hubs AMOT and EVC by differentiated links. The gene protein is reported to mediate crosstalk between the immune system and cancers, modulating cancer growth through immunosupression [169].
- ITGA3 encodes a integrin protein who's expression is linked to cancer metastasis. It is connected to FN1 by a S-link.
- Thyroglobin encoded by TG is D-linked and also found in the large connected component of the network. It is expressed at high levels in the thyroid, where it forms a substrate for synthesis of thyroid hormone.
- ITPR3 has normally ubiquitous expression in the thyroid, its protein is a regulator of calcium release receptor found in the ER. In the network is has only differentiated type links.
- ITPR1 encodes a receptor of the inositol 1,4,5-trisphosphate signal molecule, and regulates calcium release from the endoplasmatic reticulum. It has two differentiated type links, like FN1 it is connected to the high-degree gene AMOT.
- MATN2 is D-linked to FUCA1, who's expressed protein causes growth and survival of thyroid cancer cells and is involved in their invasive behaviour [144]. The other neighbor of MATN2 is SGMS1, encoding a sphingomyeling synthase also related to cancerous cell behaviour [170]. MATN2 itself encodes matrilin, a protein found in filamentous networks.
- ADGRV1 encodes a G-protein coupled receptor which binds calcium. It associates with two S-links to ANXA3 and BCL2L1. The latter is a high-degree gene encoding a protein found in the outer mitochondrial membrane where it takes part in regulation of programmed cell death pathways.
- TRIP12, Thyroid hormone receptor interactor 12, is associated directly with the hub of highest degree MICAL3 with a differentiated type link. TRIP12 found in the largest connected network component.
- HLA-DQA1 is a genes that codes for a protein of the major histocompatibility complex class II, and is central to the immune system in the recognition of antigen-presenting cells.

- S100A10 encodes a calcium binding protein which is involved with cell cycle progression and differentiation.
- NCOA4 encodes a cargo receptor mediating ferritin degradation [171]. It is D-linked to the gene PIGBOS1, which is gene that transcribes into a long non-coding RNA molecule related to cancer [172].
- FAS encodes the "Fas cell surface death receptor" protein which regulates programmed cell death. It has also been related to diseases of the immune system.
- DIO1 is linked to TPD52L1 by a differentiated link, its neighbor encodes a protein involved in cell proliferation and calcium signaling. It also positively regulates MAP3K5/ASK1-induced apoptosis.
- MAPk1 is D-linked to NDUFV2 which codes for the mitochondrial complex 1 subunit essential for mitochondrial function [173]. MAPK1 itself codes for a MAP kinase, known to be involved in regulation of cellular processes such as differentiation, proliferation of transcriptional regulation. It is normally ubiquitously expressed in thyroid glandular tissue [174].
- SLC26A4 codes for a protein called pendrin. It's function is to transport ions across the cell membrane and regulate proper balance of ions inside the cytoplasm. It has especially high basal transcription levels in the thyroid gland and inner ear. In the thyroid pendrin is related to transport iodide into thyroid glandular cells needed for thyroid hormone production. In the network SLC26A4 has degree 1 and is connected to SLC26A4-AS1 by a conserved link.
- LGALS3 has one specific link to the proto-oncogene MET, a receptor tyrosine kinase. LGALS3 encodes a galectin binding protein with numerous functions including apoptosis, immune system processes, cellular adhesion and regulation of T-cells.

## A.2 Thyroid cancer-associated genes in wTO-network

- SDHC is a gene that codes for the C sub-unit of succinate dehydrogenase, which has been investigated for relation to development of thyroid cancer [175].
- SNX19 encodes the sorting nexin-19, which is involved with intracellular vesicle transport.
- TRIP11 encodes the Thyroid hormone receptor interactor 11 is a protein-coding gene who's protein product usually associated with the Golgi apparatus and is though to be involved in it's structural organization around the centrosome.

## A.3 Thyroid cancer associated genes in MI-network

- ARRB2 expressed the gene beta-arrestin-2. This functions in regulation of G-protein receptor activity and it's expression levels is linked to several types of cancer [176].
- GSTP1 translates into a glutathione S-transferase, which mediates kinase regulation in response to cell-matrix adhesion, proliferation and inflammatory processes.
- ITGA1 codes for integrin alpha-1. This protein is involved with process regulating cellular growth.
- MLH1 DNA mismatch repair protein Mlh1 is a gene involved with the DNA mismatch repair machinery. Aberrant methylation of this gene is related to BRAF mutations in patients with thyroid carcinoma.
- PLAU codes for the urokinase-type plasminogen activator. This activator mediates cleavage of plasminogen to yield active plasmin. Together the activity of these are involved in cell proliferation and migration.
- PRKAR1A is a protein kinase-encoding gene. The gene product, type I-alpha regulatory subunit, is a cAMP-dependent regulatory unit of kinases involved in cAMP signal cascades.
- SDHB encodes for the C sub-unit succinate dehydrogenase. This is involved in the mitochondrial complex II, which is involved in the electron transport chain supplying ubiquinone.
- TCF12 as a gene that codes for the transcription factor 12, which is facilitates activation of transcription.

## A.4 Thyroid cancer associated genes in CSD-VAR-network

- CDYL expression produces the chromodomain Y-like protein. This is needed for proper regulation of epigenetic processes and acts as a co-repressor of transcription.
- PIK3CA encodes for the protein phosphoinositide-3-kinase (PI3K). It plays an important in regulation of signaling processes of cell growth, motility and morphology.

## A.5 Thyroid cancer associated genes identification chart

**Table A.1:** Summary of THCA-genes identified as differentially co-expressed by the CSD networks based on the four different similarity measurements experimented with in Chapter 5.

| Gene name | CSD | wTO | MI | CSD-VAR |
|---|---|---|---|---|
| ADGRV1 | X | | | |
| ARRB2 | | | X | |
| CCL5 | X | | X | X |
| CDYL | | | | X |
| DIO1 | X | | X | X |
| FAS | X | | | |
| FN1 | X | | X | X |
| GSTP1 | | | X | |
| HLA-DQA1 | X | | X | X |
| HLA-DRB1 | X | | | X |
| IRS1 | X | | | X |
| ITGA1 | | | X | |
| ITGA3 | X | | X | X |
| ITGB2 | X | | X | X |
| ITPR1 | X | | X | X |
| ITPR3 | X | | | X |
| LGALS3 | X | | X | |
| MAPK1 | X | | | X |
| MATN2 | X | | | X |
| MLH1 | | | X | |
| NCOA4 | X | X | X | X |
| PIK3CA | | | | X |
| PLAU | | | X | |
| PRKAR1A | | | X | |
| S100A10 | X | | X | X |
| SDHB | | | X | |
| SDHC | | X | X | X |
| SLC26A4 | X | | X | X |
| SNX19 | | X | X | |
| STAT3 | X | | | X |
| TCF12 | | | X | X |
| TG | X | | X | X |
| TPO | X | | X | X |
| TRIP11 | | X | X | X |
| TRIP12 | X | X | X | X |

# Appendix B

# Auxillary material from method development section

## B.1 Hubs in wTO-network

**Table B.1:** Hub genes in wTO-network identified as differentially expressed with degree $k \geq 40$. Predominant link type $t_{\in(C.S.D)}$, clustering coefficient $C$, and betweenness centrality $C_B$ are given for each gene respectively.

| Name | k | $t_{\in(C.S.D)}$ | C | $C_B$ |
|---|---|---|---|---|
| SP140 | 110 | S | 0.144 | 0.234 |
| TRAC | 110 | S | 0.183 | 0.181 |
| IGKV3-20 | 109 | C | 0.225 | 0.051 |
| IGHG1 | 99 | C | 0.269 | 0.022 |
| IGKV4-1 | 99 | C | 0.269 | 0.020 |
| IGLC3 | 97 | C | 0.263 | 0.016 |
| IGKV1-5 | 96 | C | 0.284 | 0.016 |
| IGKV3-11 | 90 | C | 0.315 | 0.010 |
| IGHG3 | 86 | C | 0.338 | 0.009 |
| IGKV3-15 | 79 | C | 0.382 | 0.006 |
| IGHV3-23 | 77 | C | 0.395 | 0.006 |
| IGLV2-14 | 76 | C | 0.404 | 0.005 |
| IGKV1OR2-108 | 69 | S | 0.311 | 0.088 |
| IGHA1 | 63 | C | 0.404 | 0.003 |
| IGLV1-40 | 61 | C | 0.522 | 0.002 |
| IGLV1-50 | 58 | S | 0.374 | 0.008 |
| MZB1 | 57 | C | 0.509 | 0.002 |
| XRN2 | 56 | D | 0.000 | 0.246 |
| IGHV5-51 | 56 | C | 0.575 | 0.002 |
| IGHV3-30 | 55 | C | 0.585 | 0.002 |
| PSMA3 | 51 | D | 0.000 | 0.170 |
| LSP1 | 51 | S | 0.008 | 0.203 |
| CD53 | 50 | S | 0.005 | 0.186 |
| PSMA2 | 47 | D | 0.000 | 0.115 |
| MICAL3 | 41 | D | 0.000 | 0.071 |

## B.2 GO process enrichment of all genes in the wTO-network

**Table B.2:** Enriched processes of the wTO-network

| GO biological process | FE | P-value |
|---|---|---|
| Complement activation, classical pathway | 12.39 | 1.13e-52 |
| Phagocytosis, recognition | 11.76 | 2.20e-32 |
| Regulation of complement activation | 11.49 | 7.23e-34 |
| Complement activation | 11.49 | 5.52e-52 |
| Immune response-regulating cell surface receptor signaling pathway involved in phagocytosis | 10.71 | 1.18e-35 |
| Fc-gamma receptor signaling pathway involved in phagocytosis | 10.71 | 1.18e-35 |
| Fc receptor mediated stimulatory signaling pathway | 10.56 | 2.23e-35 |
| Fc-gamma receptor signaling pathway | 10.4 | 4.17e-35 |
| immunoglobulin mediated immune response | 10 | 1.05e-47 |
| B cell mediated immunity | 9.9 | 1.91e-47 |
| regulation of humoral immune response | 9.86 | 3.02e-31 |
| B cell receptor signaling pathway | 9.82 | 1.83e-29 |
| Fc-epsilon receptor signaling pathway | 9.8 | 1.08e-39 |
| phagocytosis, engulfment | 9.62 | 1.41e-28 |
| Immunoglobulin production | 9.04 | 2.80e-30 |
| plasma membrane invagination | 8.98 | 1.75e-27 |
| Proteasomal ubiquitin-independent protein catabolic process | 8.98 | 3.97e-06 |
| membrane invagination | 8.54 | 1.11e-26 |
| production of molecular mediator of immune response | 7.92 | 5.89e-28 |
| Fc receptor signaling pathway | 7.9 | 6.45e-39 |

## B.3 Node degree distribution for weighted topological overlap-based CSD network

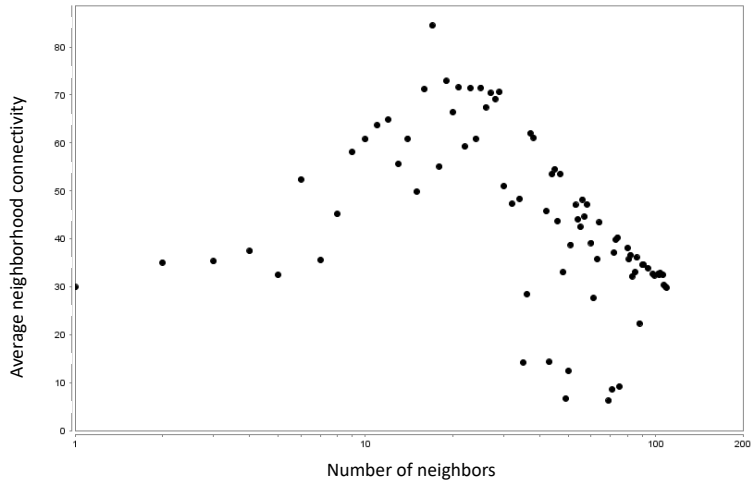


**Figure B.1:** Neighborhood connectivity as function of node degree illustration on log-log-plot for differential co-expression network inferred with CSD based on the similarity measure weighted topological overlap. Importance value of $p = 10^{-5}$.

## B.4 Hubs in MI-network

**Table B.3:** Hubs of the CSD network inferred with an importance value of $p = 10^{-5}$ based on the similarity measure mutual information. Hub genes are sorted by their respective degree $k$. Predominant link type $t_{\in(C.S.D)}$, clustering coefficient $C$, and betweenness centrality $C_B$ are given for each gene respectively.

| Name | k | $t_{\in(C.S.D)}$ | $C$ | $C_B$ |
|---|---|---|---|---|
| IGKV3-20 | 36 | C | 0.317 | 0.083 |
| IGHG1 | 36 | C | 0.317 | 0.081 |
| RPS17 | 34 | S | 0.071 | 0.143 |
| IGLV2-14 | 34 | C | 0.270 | 0.126 |
| IGKV3-11 | 33 | C | 0.352 | 0.074 |
| IGKV1-5 | 32 | C | 0.361 | 0.071 |
| IGHGP | 30 | C | 0.347 | 0.069 |
| IGLC3 | 28 | C | 0.347 | 0.045 |
| LSP1 | 27 | S | 0.043 | 0.258 |
| IGKV4-1 | 27 | C | 0.362 | 0.022 |
| IGKV3-15 | 27 | C | 0.410 | 0.038 |
| MZB1 | 24 | C | 0.522 | 0.019 |
| IGHV4-39 | 24 | C | 0.348 | 0.042 |
| IGHG3 | 24 | C | 0.4028 | 0.0334 |
| IGHV4-59 | 23 | C | 0.3878 | 0.035 |
| TRBV28 | 22 | S | 0.1088 | 0.191 |
| IGHV3-21 | 22 | C | 0.351 | 0.032 |
| IGLV1-51 | 21 | C | 0.400 | 0.030 |
| IGHV4-34 | 21 | C | 0.338 | 0.024 |
| USP34 | 20 | S | 0.084 | 0.086 |
| MFAP3 | 20 | S | 0.095 | 0.249 |
| IGLC2 | 20 | C | 0.463 | 0.014 |
| IGHV1-18 | 20 | C | 0.305 | 0.020 |
| IGHG2 | 20 | C | 0.437 | 0.015 |

## B.5 Node degree distribution for mutual information-based CSD network

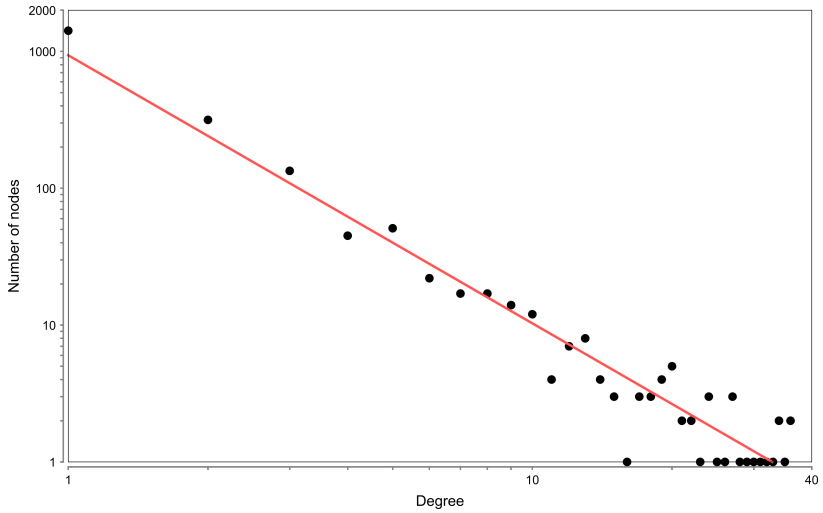


**Figure B.2:** Degree distribution plot for the CSD network based on mutual information as similarity measure. The importance level for the network is $10^{-5}$ and the axes of the plot are on a log-log-scale. The red fitted line shows the approximation of the node degree distribution with a power law function.

## B.6 THCA-associated genes from Analysis 0

**Table B.4:** Table of additional THCA-associated genes identified in network for Analysis 0. Type denotes the predominant link type among a gene's associations ($t_{\in(C,S,D)}$).

| Gene symbol | Genes Name | k | $\bar{d}$ | Type |
|---|---|---|---|---|
| ITGB2 | integrin subunit beta 2 | 9 | 3.3 | C |
| IFNAR1 | class II cytokine receptor gene | 5 | 4.1 | S |
| SDHB | Succinate Dehydrogenase Complex Iron Sulfur Subunit B | 3 | 8.0 | C |
| SNX19 | Sorting Nexin 19 | 3 | 5.3 | D |
| PMS2 | PMS1 homolog 2, mismatch repair system component | 3 | 4.6 | S |
| SMAD4 | SMAD family member 4 | 2 | 5.8 | S |
| SDHC | Succinate Dehydrogenase Complex Subunit C | 2 | 5.0 | S |
| TRIP11 | Thyroid Hormone Receptor Interactor 11 | 1 | 5.9 | S |
| MLH1 | mutL homolog 1 | 1 | 5.3 | S |
| BRAF | B-Raf proto-oncogene | 1 | 4.8 | S |
| ARNT | aryl hydrocarbon receptor nuclear translocator 2 | 1 | 1 | S |
| CDYL | Chromodomain Y Like protein coding gene | 1 | 1 | S |
| RAF1 | Raf-1 proto-oncogene | 1 | 1 | S |
| SLC26A4 | solute carrier family 26 member 4 | 1 | 1 | C |
| SDHD | Succinate Dehydrogenase Complex Subunit D | 1 | 1 | C |

## B.7 Node degree distribution for CSD network based on CSD-VAR

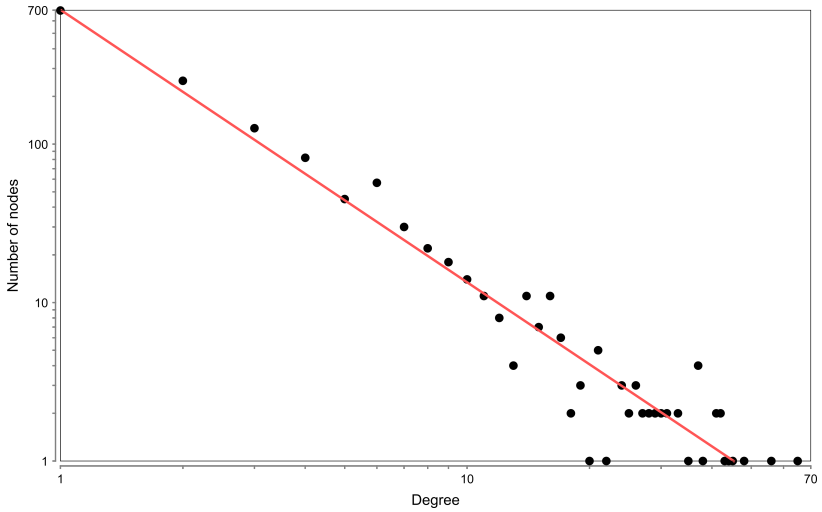


**Figure B.3:** Degree distribution plot on a log-log-scale for the network inferred with alternative Spearman's correlation coefficient not corrected for variation in correlation measures, termed CSD-VAR. The importance level for the network is $10^{-5}$. The red fitted line shows the approximation of the node degree distribution with a power law function.