Master's thesis

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

Kristin Salvesen

Using differential co-expression analysis to investigate breast cancer subtypes and the use of histologically normal cancer-adjacent tissue as the control

Master's thesis in Biotechnology
Supervisor: Eivind Almaas
Co-supervisor: Martina Hall

March 2021

NTNU
Norwegian University of
Science and Technology

Kristin Salvesen

Using differential co-expression analysis to investigate breast cancer subtypes and the use of histologically normal cancer-adjacent tissue as the control

Master’s thesis in Biotechnology
Supervisor: Eivind Almaas
Co-supervisor: Martina Hall
March 2021

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

NTNU
Norwegian University of
Science and Technology

Summary

The availability of large RNA-sequencing and DNA microarray data sets has enabled research of relationships between genes through co-expression analyses. Changes in co-expression patterns are often related to changes in biological function, and differential co-expression network analyses have become a valuable tool in the comparison of co-expression patterns between different conditions. One method for differential gene co-expression analysis is the CSD method. It compares the pair-wise correlation patterns between gene pairs from different conditions to identify conserved, specific, and differentiated associations. In this thesis an alternative CSD approach (CSD_R) was employed, using bootstrap re-sampling and existing R packages to calculate correlation and variance to reduce the computation time. This method has no filtering or selection of the resulting link scores, and the arbitrary choice of keeping the top 1000 scores of each link type was used. This selection was shown to result in a similar link distribution as calculations done with the original CSD method performed with 50 of the samples from the original data sets.

The identification of different co-expression patterns facilitates the discovery of altered interactions between different conditions and potential driving mechanisms. The CSD_R was used to analyze breast cancer-related tissues available from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), comparing them to normal breast tissue samples available from the Genotype-Tissue Expression Project (GTEx). In the first part, CSD_R was applied to histologically normal cancer-adjacent (HNCA) tissue samples extracted outside the tumor margins of breast tumors and healthy controls (HCs). HNCA breast tissue is often used as the control in breast cancer research, however, studies show HNCA as an altered intermediate state when compared to HC and breast cancer (BC) samples. Differential gene co-expression was employed with the aim of identifying changed co-expression patterns to investigate if the tumor has influences the HNCA tissue. The CSD network had many maintained interactions in processes of lipid metabolism and energy homeostasis, but pointed to a changed behavior with highly connected hubs and modules with changed interactions involved in processes like transcription and immune response.

In the second part, breast cancer tissue samples were compared to the HCs from GTEx. The aim was to identify genes and modules that could be central in breast cancer development. Breast cancer is the most commonly diagnosed cancer among woman and the leading cause of female cancer deaths. It is the second most diagnosed when combining both sexes. In the data set, the breast cancer samples were divided into five intrinsic subtypes (Luminal A, Luminal B, HER2-enriched, Basal-like, and Normal-like) with clinical and prognostic value. The CSD_R method generated networks clearly enriched in cancerous behavior, and further examination revealed highly connected hubs and modules with changed co-expression patterns linked to processes that could be involved in the underlying mechanisms of the breast cancer phenotypes. Some genes identified in the HER2-enriched and Normal-like subtype may represent novel genes involved in the development of their disease phenotype.

Sammendrag

Den økte tilgjengeligheten av store RNA-sekvensering og DNA mikromatrisestudier muliggjør forskning av sammenhenger mellom gener ved samuttrykksanalyser. Endringer i samuttrykksmønstre er ofte relatert til endringer i biologiske funksjoner, og differensielle samuttrykksnettverk har blitt et viktig verktøy for å sammenligne samuttrykksprofiler fra ulike situasjoner eller biologiske tilstander. CSD-metoden for differensiell samuttrykksanalyse sammenligner parvise korrelasjonsmønstre mellom genpar fra ulike biologiske tilstander for å identifisere konserverte, spesifikke og differensierte assosiasjoner. I denne oppgaven er det brukt en alternativ CSD-metode (CSD_R). Denne metoden bruker bootstrap for å trekke nytt utvalg, samt eksiterende R-pakker, for å beregne korrelasjon og varians, noe som reduserer beregningstiden. CSD_R-metoden har ingen utvelgelse eller filtrering av linkverdiene den beregner, og en vilkårlig grense på topp 1000 av hver linkverdi ble brukt. Denne utvelgelsen resulterte i en linkdistribusjon lignende den som ble funnet ved å bruke den originale CSD-metoden (CSD_O) med et mindre utvalg fra de originale datasettene.

Identifiseringen av differensielle samuttrykksmønstre fasiliterer oppdagelsen av endrede interaksjoner mellom ulike biologiske tilstander og mulige mekanismer som skaper den observerte forskjellen mellom to tilstander. CSD_R-metoden ble brukt til å analysere brystkreftrelaterte vev fra Molecular Taxonomy of Breast Cancer International Consortium sammenlignet med normalt brystvev hentet fra Genotype-Tissue Expression Project (GTEx). I første del ble CSD_R brukt til å sammenligne prøver fra histologisk normalt brystvev i nærheten av svulst (HNB) med prøver fra normalt brystvev (NB). HNB brukes ofte som kontroll i brystkreftforskning, men studier har vist at det er en forskjell mellom NB og HNB, og at HNB er på et eget trinn mellom NB og brystkreft. Differensiell samuttrykksanalyse ble brukt med mål om å identifisere endringer i samuttrykket for å undersøke om svulsten påvirker HNB. Nettverket viste mange bevarte, eller konserverte, interaksjoner i prosesser som lipid metabolisme og energihomeostase, men indikerte også at det var en endret adferd med nettverksnav og moduler med endrede interaksjoner involvert i transkripsjon og immunrespons.

I andre del ble CSD_R brukt til å sammenligne prøver fra brystkreftvev med prøver fra HB, der målet var å identifisere gener og moduler som kan være sentrale i brystkreftutvikling. Brystkreft er den mest diagnostiserte krefttypen hos kvinner, og den flest kvinner dør av. Det er den andre mest vanlige når man kombinerer tilfeller hos begge kjønn. I dette datasettet er brystkreftprøvene inndelt i fem undergrupper (Luminal A, Luminal B, HER2-overuttrykt, Basal-lignende, og Normal-lignende) med klinisk og prognostisk verdi. CSD_R-metoden genererte nettverk som var overrepresentert for kreftrelaterte prosesser, og videre analyse avdekket nettverksnav og moduler med endret samuttrykksmønster koblet til prosesser som kan være involvert i underliggende mekanismer som gir utvikling av kreft i de ulike undergruppene. Noen av de identifiserte genene i HER2-overuttrykt og Normal-lignende kan representere hittil ukjente gener involvert i utvikling av den gitte undergruppen.

Preface

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

I would like to express gratitude to my supervisor, professor Eivind Almaas, for his help in the form of guidance and shared insights, but also for his big enthusiasm for the field. With his support and optimism I was encourage to explore new and unfamiliar topics. My co-supervisor, Ph.D candidate Martina Hall, also deserves thanks for giving me helpful advice on analysis tools and feedback on this thesis. I would also like to thank Post.Doc. André Voigt and Ph.D candidate Jakob Peder Pettersen for their shared insights and explanations of the CSD methodology and for answering any additional questions I had.

I would also like to thank my friends for the memories and for inspiring me. These past years as a student in Trondheim would not have been the same without them. To my family who have supported and encouraged me every step of the way, and to Jonas for being a big support during the stressful time of writing a master thesis.

Kristin Salvesen
Trondheim, March 2021

Contents

List of Tables

List of Figures

Abbreviations

METABRIC	=	Molecular Taxonomy of Breast Cancer International Consortium
BC	=	Breast Cancer
ER	=	Estrogen receptor
PR	=	Progesterone receptor
HER2	=	Human epidermal growth factor receptor 2
LumA	=	Luminal A
LumB	=	Luminal B
BL	=	Basal-like
HER2+	=	Human epidermal growth factor receptor 2 enriched
NL	=	Normal-like
PAM	=	Prediction Analysis of Microarray
ROR	=	Risk of relapse
HNCA	=	Histologically normal cancer-adjacent
DEG	=	Differentially expressed gene
C, S, D	=	Conserved, Specific, Differentiated
CSD_O	=	Original CSD implementation
CSD_R	=	CSD implemented in R

Chapter 1

Introduction

1.1 Complex Biological Systems

Systems consist of interacting components and many of them interact in complex patterns collectively referred to as *complex systems* [1]. Their behavior is often difficult to predict by only looking at the components. These systems are present everywhere: Cities connected by roads, airports connected by flights and social interactions either by physical meetings, calling or social media interactions. These are just a few of the complex systems that are a part of our everyday life.

In a desire to understand the behavior of the complex systems around us the field of *network science* emerged [1]. This field aims to describe and understand the intricate network of interacting components, and ultimately to predict and control the future behavior of such systems. It builds upon graph theory, and utilizes both mathematical descriptions and computational modeling to obtain understanding of complex systems and their properties. Any system can be represented as these networks and understanding them may lead to improvements in anything ranging from logistics and communications to understanding disease mechanisms and medicine development.

The complex systems of biology are many, ranging from the Amazon rainforest, full of animals, insects and plants that live together, compete with each other and depend on each other, to the energy production of the mitochondria in a cell. This microscopic system is only one of many systems found in the human body, which in itself is a world of many systems. A world consisting of cells, tissues and organs with specialized and intricate functions, and a collaboration between them that is vital to maintain the body. The sub-field of network science analyzing such biological networks on a molecular level is *systems biology* [2]. This include biological networks from other animals, plants and microorganisms, as well as humans. Two types of networks, protein-protein interaction networks and gene co-expression networks, are commonly studied in this field [3, 4, 5]. In a protein interaction network the nodes represent proteins and the links between them represent physical interaction, while in a gene co-expression network the nodes represent genes and the link between them represent the correlation of their gene expression, that primarily function to synthesize a given protein.

As in any system things can go wrong in the cellular systems of our body. Generally these are harmless and have few consequences, but sometimes the changes can have a major impact affecting our health and life. When the changes result in disease, the field of medicine try to find a solution to counter the given disease. In this venture it is essential to have knowledge about the underlying mechanisms and causes of the disease phenotype.

To investigate the cause of the diseased state a reasonable starting point is the genes and proteins. As varied as they are, all biological systems are build up by the same building blocks. The cells of our body contain the same DNA, which describe their appearance and functioning through a blueprint of genes. Despite of this the cells can have a variety of functions and structures, given by the collection of genes they express. This collection is different for different cell types and result in diverse tissues and functions. The genes contains the recipe of RNA, which in turn is translated into proteins. Any cell constantly monitors the internal and external environment as an input to handle and respond to changes in the environment as well as possible, by mechanisms to repress or activate the gene expression [6]. Mistakes in this highly regulated system may therefore result in change in the cell's behavior.

Disruptions in these complex systems, known as mutations, can lead to diseases. Some diseases is caused by a single gene mutation [7], but most are a result of multiple mutations in several genes. Techniques like RNA microarrays and RNA sequencing allow measurement of the gene expression in cells at a given time and can be used to investigate changes in the gene expression between different conditions, for instance between healthy and diseased states. Analyzing different gene relationships and networks can give researchers insight into genes that contribute to underlying mechanisms of the disease phenotype.

One of the approaches to understand diseases is gene co-expression analysis methods, in which the relationship between genes is investigated by *co-expression*, i.e. how correlated their gene expression is. Another approach it the *differential* gene co-expression analysis, in which the co-expression pattern of gene pairs from two or more conditions are compared. These analyses can be used to a variety of comparisons, for instance between tissue types, species, treatment and control, diseased and healthy tissues. The revealed gene co-expression patterns often relate to biological functions [8, 9, 10], and present a powerful tool for identifying mechanisms or processes behind for instance cell differentiation, disease progression and cell responses to treatments or drugs. Using this to study diseases can potentially give better understanding of the dysfunctions behind a given diseased phenotype and point to genes of interest for treatment or drug developments. One method for differential gene co-expression analysis is the CSD method [11], which identifies conserved (C), specific (S) or differentiated (D) gene co-expression patterns of gene pairs with strong correlations between two conditions. Data sets of gene expression measurement from different conditions, tissues and species are now freely available from multiple online databases and can be used for these analyses [12, 13].

A group of diseases that can be investigated by these analyses is cancer, which is the result of an accumulation of mutations. Here, the regulation of cell growth and cell death is disturbed, transgressing the most basic rules of cell behavior in a multicellular organism [14, 15]. This process is a gradual development, and multistep accumulation of biological capabilities, enabling tumorigenesis and ultimately malignancy. The end result is a growing tumor and potentially its metastatic propagation, which caused an estimated 18.1 million new cancer cases and 9.6 million cancer deaths in 2018 [16].

The most prevalent type of cancer when combining both sexes is lung cancer (11.6%), closely followed by breast cancer (11.6%), with a difference of about 5,000 cases [16]. When only looking at the female population, breast cancer is the most commonly diagnosed cancer type and the leading cause of cancer death.

Breast cancer is a complex disease, but also a heterogeneous disease, with identified intrinsic subtypes distinguishing between tumor types displaying separate behaviors, prognosis and gene expression patterns. Perou et al. divided breast cancer into five intrinsic subtypes based on their genetic characteristics [17, 18]. Almost a decade later Parker et al. created a 50-gene classifier for these subtypes, the PAM50-classifier, assigning each tumor sample into one of the subtypes and estimating a risk of recurrence score (ROR) [19]. This is clinically used as the Prosigna test to guide decision-making on adjuvant systemic therapy in certain tumor types [20, 21, 22].

The large number of genes involved in controlling cell growth, death, and differentiation highlight the importance of analyzing the genetic expression patterns of thousands of genes in concert to investigate their correlation and behavior. This is often done by using histologically normal cancer-adjacent tissue samples extracted outside the tumor margins as the control, assuming that normal histology implies biological normalcy. However, research has shown a difference between histologically normal cancer-adjacent tissue and breast tissue without a tumor present [23, 24, 25]. This presents a problem in using these samples as controls, and breast tissue without tumor present has been shown to identify additional differentially expressed genes [23].

1.2 Aims and objectives of this thesis

This thesis has two main goals that are closely related. For both of them the CSD-framework is the primary tool for analysis, supplemented by enrichment and network analysis tools. The first goal is to perform a differential gene co-expression analysis focusing on the changes in histologically normal cancer-adjacent (HNCA) tissue in comparison with breast tissue without tumor present. This is motivated by studies showing changes in HNCA tissue compared to normal breast tissue, and the current use of HNCA tissue as control in breast cancer research. This is done by using gene expression measurements of HNCA samples from METABRIC and comparing them to normal breast tissue expression profiles from the GTEx project, in order to perform an in-dept investigating looking at how a tumor may influence the benign-looking surrounding tissue.

The second goal is to identify transcriptional alterations in well-established molecular subtypes of breast cancer to investigate genes and modules relevant for each subtype. This is done by comparing each of the breast cancer subtypes with the healthy controls of breast tissue from the GTEx project. The identification of network modules and relevant genes is done by network analysis. The identified modules could represent disease modules that potentially could reveal novel patterns and genes that contribute to the underlying mechanisms resulting in the cancerous phenotype.

These two research goals are reflected in the presented work and the following sections is organized accordingly. Summarized the aims of this thesis are to:

1. Perform a differential gene co-expressed analysis comparing breast tissue samples from healthy individuals and tissue samples taken adjacent to breast tumor, with normal histology, in order to investigate relevant modules of genes and explore the effect of using histologically normal cancer-adjacent tissue samples as the control in breast cancer research.

2. Perform a differential gene co-expressed analysis comparing tissues from different breast cancer subtypes with breast tissue samples from healthy individuals to investigate relevant genes and modules that may contribute to breast cancer development and progression.

Chapter 2

Background

This chapter will introduce the main topics and the theoretical founding of the methods used in the analyses in this thesis. As some of these topics are vast, the following sections aim to provide background information underlying the methods, results and discussion.

Theory about systems and network biology is obtained from *Network science* by Albert Lázló-Barabási [1] and *A first course in systems biology* by Eberhard O. Voit [2] unless other sources are stated. The reader is referred to them for more detailed and extensive information. The CSD method for differential gene-co expression developed and described by Voigt et al. [11] is the source for the information provided in Section 2.8.

2.1 Breast Cancer

Breast cancer (BC) is the most commonly diagnosed cancer among women and the second most commonly diagnosed for both sexes combined [16]. It is the leading cause of cancer death among women, with an incidence rate far exceeding other cancers regardless of HDI (Human Development Index). About 5% to 10% of breast cancer cases is accounted for by hereditary and genetic factors, including history of breast or ovarian cancer and inherited mutations in breast cancer susceptibility genes, such as *BRCA1* and *BRCA2*.

Breast cancer is a heterogeneous disease, not only on a molecular level, but also the cellular composition and clinical outcome [19]. Availability of gene expression profiles solidified the notion of molecular characteristics influencing prognosis and treatment response, complementary to clinicopathalogical parameters [26]. This has evolved the treatment concepts, aiming at more biologically driven therapies accompanied by the traditional clinicopathalogical parameters, such as tumor grade (differentiation) and biomarker receptor status [27], when making treatment decisions [28]. Receptor status is a treatment predictive factor and breast cancers are routinely scored for oestrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor 2 (HER2) protein expression by immunohistochemistry (IHC) [27].

Several classifications have been developed to identify molecular alterations that can be used for prognosis and to help guide treatment decisions, such as the 21-gene Oncotype

DX assay, the 70-gene MammaPrint microarray assay [22]. In 2000, Perou, Sorlie and colleagues reported an intrinsic classification, distinguishing four breast cancer subtypes [17]. Later, these where expanded by dividing one subtype into two, resulting in five subtypes: Luminal A (LumA), Luminal B (LumB), Basal-like (BL), Normal-like (NL) and Human epidermal growth factor receptor 2 enriched (HER2+) with prognostic impact [18]. This classification shifted breast cancer management from being based on tumor burden to biology-focused approaches [28]. Their characteristics are summarized in Table 2.1.

Table 2.1: Molecular status of the clinical markers oestrogen receptor/progesterone receptor/human epidermal growth factor 2 (ER/PR/HER2), characteristic genes, and general characteristics of each of the intrinsic subtypes [26, 27, 29].

Subtype	ER/PR/HER2	Characteristic genes	Characteristics
Luminal A	ER+ PR+/- HER2-	*ESR1, KRT8, KRT18, GATA3, XBP1, FOXA1, TFF3, CCND1, LIV1*	Most common (40%-70%) Best prognosis Low proliferation gene cluster expr. Low histological grade Low occurence of *TP53* mutations
Luminal B	ER+, PR+/-, HER2-/+	*ESR1, KRT8, KERT18, GATA3, XBP1, FOXA1, TFF3, SQLE, LAPTM4B*	10%-20% Worse prognosis than Luminal A Higher proliferation gene cluster expr. Intermediate histological grade *TP53* mutations
HER2-enriched	ER- PR- HER2+	*ERBB2, GRB7*	5%-15% Aggressive High proliferation gene cluster expr. High histological grade *TP53* mutations
Basal-like	ER- PR- HER2-	*KRT5, KRT17, CDH3, FABP7, TRIM29, LAMC2, ID4, EGFR,*	15%-20% Worst prognosis High proliferation gene cluster expr. High histological grade *TP53* mutations
Normal-like	ER+/- PR+/- HER2-	*PTN, CD36, FABP4, AQP7, ITGA7*	Rare Good prognosis Low proliferation gene cluster expr. Low histological grade Low occurence of *TP53* mutations

The intrinsic subtypes were classified by "intrinsic" gene lists including genes with significantly larger variation between different tumors than between samples from the same tumor, representing inherent properties of the tumor itself [17, 18]. The gene lists were used for hierarchical clustering, resulting in the classification of the intrinsic subtypes and clusters of genes for the identified subtypes. Although effective in identifying the subtypes, the method is not suitable for single sample classification and clinical use, as identification of one new sample require reanalysis of all samples. Investigating an unchanging and objective classification, Sørlie et al. computed centroids (mean expression profiles of the intrinsic gene list) for each of the subtypes including only the tumor samples with the highest correlation within each subtype, using prediction analysis of microarrays (PAM)

[30]. This is a nearest-centroid classification with an automatic gene selection step integrated into the algorithm, to obtain centroid prediction from a minimal number of genes. PAM increasingly shrink the centroids by a shrinkage parameter Δ from no shrinkage to complete shrinkage, and identifies a minimal set of genes that predict the centroids/subtype accurately [31]. This yielded a strong agreement (>79%) between the hierarchical clustering and the PAM predictions of different data sets [30].

2.1.1 The PAM50 Subtype Classifier

Almost a decade after the initial intrinsic subtypes Parker et al. developed a 50-gene classifier of the intrinsic subtypes and a risk of relapse (ROR) score, using the PAM algorithm for centroid construction [19]. The gene list, hence referred to as PAM50, and their relative expression in each of the subtypes are available in Appendix A.1. It provides additional prognostic and predictive information to standard parameters for breast cancer patients. Furthermore, the ROR score is valuable for management of breast cancer that has not spread to the lymph nodes (node-negative). In this classifier the normal-like subtype was represented with normal tissue, and thus NL is considered a quality-control measure and not included in outcome analyses or calculation of ROR score.

The subtype classification revealed that close to 10% of tumor samples were normal-like, and as this was developed by normal breast tissue samples, Parker et al. speculate in the class being an artifact of tumor specimen with normal contamination. The normal-like group in the two initial subtyping also included normal breast samples [17, 18]. However, other research point to it being a genuine subtype [32].

Initial classification of the intrinsic subtypes classified an initial branching based on ER status in the hierarchical clustering; the ER+ branch with LumA and LumB, and the ER- branch with HER2+, BL and NL [18]. This clinical marker status was confirmed for the majority of samples within each subtype by Parker et al., although all subtypes were represented in ER+, ER-, HER2+ and HER2- categories, demonstrating that clinical marker status alone is not adequate in identifying the intrinsic subtype of a tumor [19]. This is further corroborated by Bastien et al. comparing PAM50 subtyping with a surrogate subtyping using IHC markers ER, PR and HER2 [33].

Clinical trial constitute a Level 1 evidence for clinical validity of the PAM50 test in predicting the risk of distal recurrence (DR) in postmenopausal women with ER+ early breast cancer [34]. Discriminating between low- and high-risk groups that would be unlikely and likely, respectively, to benefit from additional chemotherapy to improve the outcome. Currently, the PAM50 classifier is available as the Prosigna test, and is recognized as valuable for clinical use by several guidelines [20, 21, 22]. The American Joint Committee on Cancer (AJCC) eight edition staging manual include Prosigna as a stage modifier for hormone positive, HER-, lymph node negative (H+, HER+, LN-) patients scored with a low ROR score, placing the tumor at a lower stage regardless of tumor size. The American Society of Clinical Oncology (ASCO) clinical practice guidelines recommend the ROR score in guiding decision-making on adjuvant systemic therapy in H+, HER+, LN- tumors, while the European Group on Tumor Markers (EGTM) also include lymph-node positive patients.

2.1.2 Histologically Normal Cancer-Adjacent Breast Tissue

The tissue in the regions immediately surrounding the tumor have morphological and phenotypic changes distinctive from healthy tissue without a tumor present, for instance pH levels, and transcriptomic and epigenetic aberrations [23]. These are apparent up to 1 cm from the tumor margins, and consequently, histologically normal cancer-adjacent (HNCA) samples are taken adjacent to the tumor but beyond these observed changes. The HNCA samples are often used as control samples for cancer research with the assumption that normal histology implies biological normalcy. Such tissue samples are readily available from reduction mammoplasty and prophylactic mastectomy. However, little is known about how HNCA tissue is influenced by the tumor or how its expression profile compare to tissue from non-diseased individuals.

Ever since the theory of "field cancerization" suggested a cumulative, step-wise process of obtaining genetic alterations in carcinogenesis, leaving molecular alterations in morphological normal adjacent tissue [35], the "normalcy" of HNCA tissue has been debated. Studies have shown a difference between HNCA and breast tissue without tumor present (hereby referred to as healthy tissue) [23, 24, 25]. These studies point to HNCA tissue reflecting the intrinsic subtype, in an intermediate, distinct state between healthy tissue and tumor, with activation of pro-inflammatory response genes. Genes identified to relate to molecular alterations in HNCA tissue from different tissue types are available in A.2.

Using HNCA samples as the control for differential expression analysis in cancer studies have been shown to identify the majority of differentially expressed genes (DEGs), although using healthy tissue provides additional information and may reveal obscured biomarker candidates or therapeutic targets [23]. In this study the tumors vs. healthy analysis found more significant DEGs than in tumor vs. HNCA and a discordance between up- and down-regulation in 93 breast cancer genes.

2.2 Gene Expression

Even though every cell in the human body has the same genetic material, cell types differ greatly in function and composition, i.e. a neuron cell differ greatly from a skin cell. A fundamental differentiation appear during development as various signals or growth factors guide cells to differentiate into the different cell types, by changing their gene expression. After settling into their specific cell type, the cell continues a *differential* gene expression in response to signals. The cells continuously monitors the internal and external environment for these signals, such as nutrient availability, signal molecules from neighboring cells and damage, to produce the appropriate proteins in response. The response is aided by up- or down-regulation of genes encoding transcription factors (TFs). TF proteins regulates the gene expression further, enabling the cell to quickly handle and respond to the current environment [6].

TFs regulate the gene expression by physically binding to the DNA, affecting the binding affinity of RNA polymerase and aid or prevent access to a certain part of the DNA. Consequently, activating or repressing transcription of genes in this region, that are either TFs themselves or have a different function in the cell [6]. The active genes that are transcribed into their corresponding messenger RNAs (mRNAs), which are subsequently processed an translated to proteins, the final gene product. The full set of mRNA (and other RNAs) transcripts in a cell at a specific time, and their quantity, is the *transcriptome* of the cell [36].

2.2.1 RNA Sequencing

To capture the transcriptome, RNA sequencing (RNA-seq) is a widely used tool. RNA-seq quantifies and identifies the RNA transcripts present in a biological sample at a given time, providing a transcriptome profile. Here, RNA is isolated, the transcripts are converted to complementary pieces of DNA (cDNA), followed by high-throughput sequencing methods used to align and identify the transcripts, as well as the relative abundance of each transcript [36]. The genes expressed at the given time in a certain tissue is the *expression profile* of the sample, and reveal the activity of the genes [37].

Gene expression profiles can be obtained during different conditions, for instance in different developmental stages or in response to disease or treatment, making RNA-seq a powerful tool to investigate the responses to change [36]. A wide variety of RNA-seq data from many different conditions, tissues, and species are available from different publicly available online databases, such as the Gene Expression Omnibus (GEO, [12]) and the Genotype-Tissue Expression Project (GTEx, [13]).

Differential gene expression analysis can be performed on these data sets to identify genes that are expressed differently across two or more conditions or in response to some factors of interest [38]. Performing RNA-seq on tissue samples from different conditions of interest provides the specific expression profiles for each condition, and can be compared to investigate the cell's response to changes. This is done by statistical comparison between the conditions (see Section 2.4). Such analyses can provide insight into genes that may contribute to underlying mechanisms of the cell's response to the given condition or of the disease phenotype, for instance a mechanism underlying a disease. Potentially resulting in new treatment methods by identifying targets for new drugs [38].

2.3 Network theory

A network is defined as a collection of nodes connected with links. This can be used to describe naturally occurring or man-made systems, where the components of the system (*nodes*) interact (*links* or *edges*). This general set-up makes it possible to define and study a network for any system consisting of components that interact or have some kind of relationship or connection, e.g.cities connected by roads, social interactions or neurological signaling in the body. Once the network is defined it can easily be visualized. One options is to represent nodes as circles and links as the lines connecting them, as illustrated in Figure 2.1.

2.3.1 Connectivity and Adjacency Matrix

The number of nodes N in a network denotes the size of the network. Any given node *i* in the network is connected to a number of other nodes. These other nodes are node *i*'s

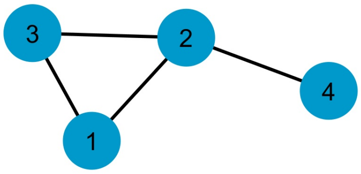


Figure 2.1: Visualization of a network consisting of four nodes represented by blue circles and the links between them represented by black lines. The nodes could e.g. represent computers or people and the edges represent information exchange between the computers or a social relationship between people.

nearest neighbors, and the number of nearest neighbors is equal to the *degree*, k_i, of the node.

The most common mathematical representation of a network is in terms of the *adjacency matrix* A = [a$_{ij}$], where

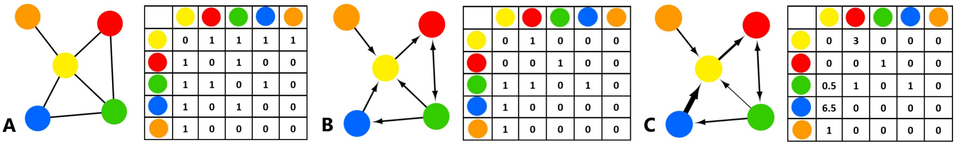
$$a_{ij} = \begin{cases} 1 & \text{if there is a link } \textit{from} \text{ node } i \text{ to node } j \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

This mathematical representation can handle many types of networks. Figure 2.2 shows three types of networks: (A) undirected and unweighted, (B) directed and unweighted, and (C) directed and weighted, and their corresponding adjacency matrices. Note the unique correspondence between the adjacency matrix and how the nodes are connected by links.

In an *undirected* network, the link does not have a direction and the interaction is equal, i.e the link carries no additional information. This adjacency matrix is symmetrical and binary, see Figure 2.2A, with a$_{ij}$ = a$_{ji}$. If the connection between nodes are directional, e.g. A affects B, but B does not affect A, the network should then include links with a designated direction according to the interaction, making it a *directed* network. Resulting in an asymmetric adjacency matrix: a$_{ij}$ $\neq$ a$_{ji}$. *Weighted* networks quantifies uneven importance or value of links, e.g. signal intensity, by assigning a weight to the link where the value is given by ω_{ij}. The adjacency matrix entries then contain a continuous range of numbers reflecting this property. Such weighted networks can be converted to unweighted networks by defining a threshold value for the weight, keeping all links with a weight above this threshold and discarding those below.

Every connection to a node *i* is represented by the *i*'th row and the *i*'th column in the adjacency matrix. In an undirected network the value and order of the elements in the *i*'th row and *i*'th column is equal, with matrix symmetry around its diagonal. In a directed network they may differ as each of them represent one direction of the interaction: *i*'th row from *i*, and *i*'th to *i*.

The degree of the nodes, i.e. the number of connections, can be found in the matrix. For undirected networks the degree can be found by counting the non-zero entries, either in the *i*'th row or column, in the adjacency matrix, while for a directed network the *total degree* of node *i* (sum of its *in-degree*, k_{in}, and *out-degree*, k_{out}) can be found by counting the non-zero entries of both the *i*'th row and *i*'th column. The entries in the diagonal of the adjacency matrix is zero, unless the node interact with itself to make a self-link.



	🟡	🔴	🟢	🔵	🟠
🟡	0	1	1	1	1
🔴	1	0	1	0	0
🟢	1	1	0	1	0
🔵	1	0	1	0	0
🟠	1	0	0	0	0

	🟡	🔴	🟢	🔵	🟠
🟡	0	1	0	0	0
🔴	0	0	1	0	0
🟢	1	1	0	1	0
🔵	1	0	0	0	0
🟠	1	0	0	0	0

	🟡	🔴	🟢	🔵	🟠
🟡	0	3	0	0	0
🔴	0	0	1	0	0
🟢	0.5	1	0	1	0
🔵	6.5	0	0	0	0
🟠	1	0	0	0	0

Figure 2.2: Three networks and their corresponding adjacency matrix. A) Undirected unweighted network, B) directed, unweighted network, C) directed, weighted network.

A *connected component* of an undirected network is a set of nodes connected so that it is possible to start from any node in the component and, by following the links of the network, reach any other node in the component. If a connected component of the network is much bigger than any other components in the network it is referred to as the *giant component*.

2.3.2 Degree Distribution and Scale-free Networks

The *degree distribution* is a network property referring to the proportion of nodes in a given network that has the degree *k*. In a random, artificial network, such as an Erdös-Rényi network, links are randomly associated with nodes and the degree distribution is binomial, resembling a bell curve with small variance. Resulting in most nodes having a degree that is close to the average degree.

In contrast to the binomial distribution within random networks, biological and other real-world systems are often observed with a power-law distribution, characterized by being so-called *scale-free* networks. This is formulated in Equation 2.2, where p_k is the probability that a node will have the degree *k*, and γ is the degree exponent:

$$p_k \sim k^{-\gamma} \tag{2.2}$$

A scale-free network is characterized by a few highly connected nodes and many nodes connected to only a few other nodes. Nodes with disproportionately more links than the average node are called *hubs*. A feature of scale-free, hub-containing networks is that the the shortest path length, or *distance*, between two randomly selected nodes are noticeably shorter than in a random network. The path length is defined as the number of links needed to go from one node to another. In scale-free networks these paths often go through hubs minimizing the distance, as they provide a path intersection for many non-hub nodes. The hubs contribute to a robustness in scale-free networks against random attacks, suck as removing a node or changing a link, as there are many alternative routes through the hubs. However, this structure also makes the network more vulnerable to targeted attacks against these central hubs. Selectively removing only a few hubs sufficiently breaks down the network, partially or fully depending on the number of removed hubs.

2.3.3 Assortative and Disassortative

Assortative networks are defined by the tendency of nodes of similar degree to connect to each other, hubs connecting to hubs and small-degree nodes connecting to small-degree nodes. *Disassortative* networks, on the other hand, have the tendency of similar nodes to avoid linking to each other in the network, and instead hubs and small-degree nodes connect to each other.

This correlation can be detected by inspecting the *neighborhood connectivity distribution*, plotting the average connectivity (degree) of nearest neighbors of nodes with degree k as a function of degree k itself. An approximated line through the points gives the degree correlation function, $k_{nn}(k)$, expressed in Equation 2.3:

$$k_{nn}(k) = ak^{\mu} \tag{2.3}$$

where k is the degree, μ is the correlation exponent, and a a regression constant. Here, an increasing $k_{nn}(k)$ with k indicates an assortative network, as high-degree nodes tend to link with other high-degree nodes. In the opposite case, a decreasing function indicates that the network is disassortative.

Consequently, the degree correlations of a given network is dependent on the correlation exponent, μ, as follows:

- Assortative network: $\mu > 0$
- Neutral network: $\mu = 0$
- Disassortative network: $\mu < 0$

These degree distributions for a network are illustrated in Figure 2.3: one assortative, one neutral and one disassortative.

2.3.4 Node Parameters

A previously discussed node parameter is the node degree, k, which is an important parameter of scale free networks as it tells how connected the node is to the rest of the network. Another important characteristic is the centrality of a node, or the placement. Nodes with the same degree can be on the outskirts of the network or closer to the center. This can be measured by different centrality measures like the *betweenness centrality* and *closeness centrality*. The betweenness centrality describes the number of shortest paths passing through the node, while the closeness centrality is the sum of shortest paths from that node to all other nodes in the connected component. Another centrality measure is the *eccentricity* of a node, defined as the longest path to any other node among the shortest distances.

2.3.5 Network Parameters

Networks can be characterized by several different parameters that describe different networks properties. The previously discussed degree distribution is one of them. The *diameter* of the network is the largest of the shortest distance between all node pairs, i.e.

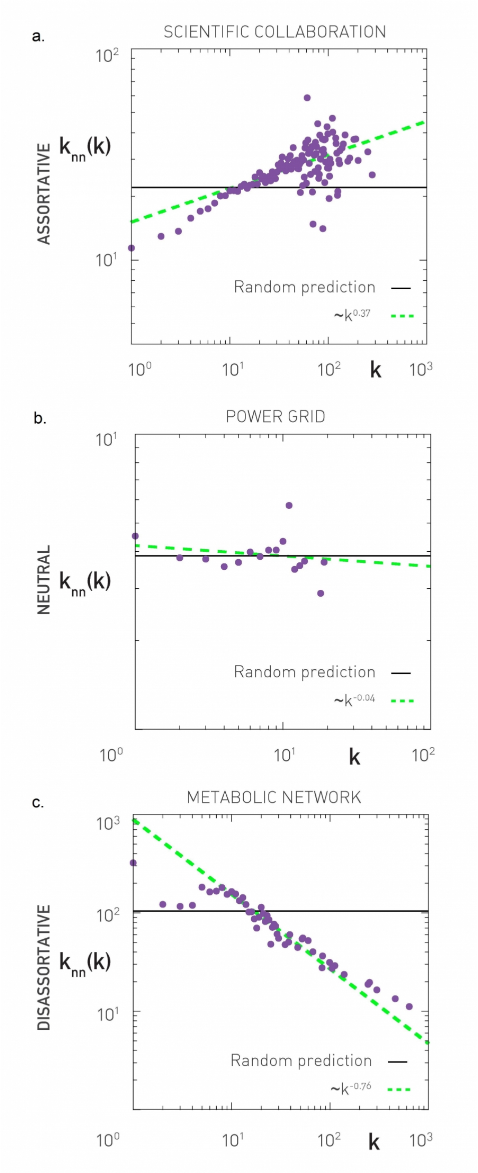


Figure 2.3: Neighborhood connectivity distribution for three real networks. The degree correlation function $k_{nn}(k)$ on a loglog plot for a collaboration network, a power grid, and a metabolic network, showing assortative ($\mu = 0.37$), neutral ($\mu = 0.04$) and disassortative networks ($\mu = 0.76$), respectively. In each plot, the green, dotted line giving the regression line through the points, and the black, horizontal line corresponds to the expected degree correlation if it was completely random. Source: [1]

the largest node eccentricity, while the *radius* is the smallest non-zero eccentricity. The average distance between the nodes in a network defines the *characteristic path length*.

While the network itself is scale-free, local sub-networks within the network may have a different structure and properties than the whole network.

Modules

A module is a subnetwork of nodes more closely related to each other than the rest of the components in the network having a high *clustering coefficient*, C_i [39]. The clustering coefficient describes to what extent the neighbors of a given node *i* connect to each other expressed by

$$C_i = \frac{2L_i}{k_i\,(k_i - 1)} \tag{2.4}$$

where L_i denotes the links between the k_i neighbors of node *i*.

This method identifies structural modules in terms of the topology and has been useful to identify functional components in a network [40]. The identified clusters can be enriched for genes cooperating in a specific biological function, making them not only structural modules but also functional modules, which again can result in a disease module if a breakdown occurs. Note that it is likely that the disease module is not identical to the functional/topological module, but more probable that it overlaps with it [41].

Topological Overlap

Topological Overlap (TO) look for information in the neighborhood of a gene pair, i.e. third party genes that are connected to that gene pair [42]. The topological overlap reveal communities within the network and reflect their relative interconnectivity. The score is given by the fraction of nearest neighbors *shared* by both node *i* and *j*:

$$\omega_{ij} = \frac{\sum_k a_{ik}a_{kj} + a_{ij}}{min\{k_i, k_j\} + 1 - a_{ij}} \tag{2.5}$$

where $\sum_k a_{ik}a_{kj}$ is the sum of shared neighbors between node *i* and node *j*, a_{ij} = 1 when there is a interaction between node *i* and *j*, and k_i and k_j is the node degree of node *i* and node *j* respectively [40]. A topological overlap of 1 indicate that node *i* and *j* is connected to the same neighbors and a score of 0 that they share no links to their neighbors. Giving links with visible clusters a high score.

For weighted networks, links contain a strength value ω_{ij}, and the weighted topological overlap (wTO) can be calculated by

$$\omega_{ij}^{wTO} = \frac{\sum_k \omega_{ik}\omega_{kj}a_{ik}a_{kj} + \omega_{ij}a_{ij}}{min\{s_i, s_j\} + 1 - |\omega_{ij}a_{ij}|} \tag{2.6}$$

including the weights and substituting the node degree *k* with the node strength $s_i = \sum_j a_{ij}|\omega_{ij}|$ [43].

2.4 Statistics

2.4.1 Correlation

When constructing a network, one needs to systematically decide whether or not there is a link between two nodes, in order to capture meaningful interactions. The link between two nodes often reflect a continuous measure, like the correlation value for a given property, rather than a binomial value. This is also the case for for gene co-expression networks. Any link in these networks represent a correlation of the RNA-levels between two genes.

This correlation is measured by the correlation coefficient, often denoted ρ, and represents the strength of the linear relationship between two variables [44]. The correlation coefficient takes any value from -1 to 1. The closer ρ comes to ±1 the stronger the correlation is, positive or negative. The value zero indicate an absence of a linear relationship. If the coefficient is positive, there is a positive relation reflecting that the variables vary in the same direction. A negative correlation, denoted by a negative coefficient, reflects an inverse association where a a high score in one variable is accompanied with a low score in the other, and vice versa. The closer ρ is to zero, the less do the two variables follow a given association or relationship.

The correlation coefficients observed between the expression of two genes is often not zero, though they are not necessarily high either. To create a meaningful network, it is not enough to only identify any relationship between the gene expression, but to find the strong and systematic relevant relationships between them. Therefore, it is important to separate correlations that could easily happen by chance from those that are unlikely to happen unless there is a mechanism behind them.

Pearson Correlation

The Pearson correlation coefficient measure the strength of linear relations. It can be used to describe linear associations between two variables or two series of data measurements with joint distributions, such as normalized gene expression data sets from RNA-seq. The Pearson correlation coefficient, ρ_{ij}, is given by Equation 2.7 [44].

$$\rho_{ij} = \frac{cov(i,j)}{\sigma_i \sigma_j} \tag{2.7}$$

Here *cov(i,j)* denotes the covariance between *i* and *j*, σ_i and σ_j the standard deviations of *i* and *j* respectively.

Spearman Correlation

The Spearman correlation is of the same form as the Pearson correlation, except that the ranks of the data series replaces the observed values in the correlation calculations [45]. This ranking is done by assigning the smallest value a ranking of 1, and subsequently, with an increment of one, assigning increasing values their rank. The rank replace the measured value and represent the relative value of the measurement compared to other measurements in the same sample. For example, a data series with the values 2.1, 4.6, 3.0, 1.5 and 2.4 would have the rank 2, 5, 4, 1 and 3 respectively. This makes it possible to compare correlations from two different data sets that are not normalized or normalized differently.

Before calculating the Spearman correlation of a gene pair, the gene expressions in each sample of the data sets has to be replaced by their rank. Following this, the Spearman correlation, $rho_{r(i)r(j)}$, given in Equation 2.8, can be calculated.

$$\rho_{r(i)r(j)} = \frac{cov(r(i), r(j))}{\sigma_{r(i)}\sigma_{r(j)}} \tag{2.8}$$

where *r(i)* and *r(j)* is the ranks of gene *i* and *j* respectively, *cov(r(i),r(j))* is the covariance between the ranks of *i* and *j*, and $\sigma_{r(i)}$ and $\sigma_{r(j)}$ is the standard deviations of *r(i)* and *r(j)*, respectively.

2.4.2 Bootstrapping

Bootstrapping is re-sampling of a set of data and making statistical calculations of this sample. Bootstrap uses random sampling with replacement, where a sample is drawn from a finite amount of samples and returned to the sample pool before the next unit is drawn [46]. Every re-sampled set is of the same size as the original set. For example, for a data set with 20 samples a bootstrap iteration draws 20 samples, resulting in a re-sample that may contain duplicates. This re-sampling (bootstrapping) is thought to model the unknown population, as the distribution of the samples taken from that population can be a guide to the distribution of the parent population.

The bootstrap is a computer-intensive method repeated multiple times and statistical conclusion are made from the resulting re-sampled sets [46]. Following a re-sample, the statistics of interest is calculated for this sample collection [47]. When all bootstrapping iterations and statistical calculations for each re-sample, the mean of the calculated statistics is computed.

Running the bootstrap with too few re-samplings, or iterations, can give varying results for each bootstrap analysis, resulting in both significant and insignificant results from the same data [48]. Consequently, resulting in unreliable results. A commonly used approach to determine the number of iterations needed is to start with a given number of iterations, for instance 100 bootstrap re-samplings, and then double it to 200 to compare the approximations [47]. This is then repeatedly increased until the observed change is small enough, i.e. the compared results are consistent.

2.4.3 Confounding

Confounding is the situation when an evident correlation between two observations is caused by a third factor, correlation with one or both of the observed variables [49, 50]. The additional, hidden factor(s) not accounted for that cause or distort the relationship between two variables are referred to as the *confounding factor*. This can easily lead to false conclusions of direct relationship, when there in fact is a spurious correlation. General characteristics of a confounding factor include that it is predictive of the outcome in the absence of the exposure, it is associated with one or both variables, and it is not an intermediate between exposure and outcome. An example could be the following fictional study of the health between a group of people training and a group of people that doesn't. When concluding that the people training have a better health, there is a possibility that different diets is an alternative or contributing explanation. Making diet a confounding factor.

Even if one carefully try to avoid obvious sources of confounding, it is difficult to remove, control or measure all of the possible confounding factors in an experiment. If a gene pair in a data set shows a significant co-expression in a given tissue type and is consistent across all samples, it is likely that the co-expression pattern between these two genes is typical for this condition. On the other hand, if a more detailed review show that certain subgroups have a very high co-expression, while this strong co-expression is not present outside this subgroup, the overall correlation coefficient could still be significant. In this case, there is a risk that this significance is due to confounding factor(s) these have in common, such as the age, an unreported disease, lifestyle etc. of the individuals the samples were collected from, and not due to the given condition itself. Consequently, confounding factors tend to result in a high variance in the correlation calculated from different subset of the full data set.

2.4.4 Hypothesis Testing

Hypothesis testing is a statistical procedure comparing two hypotheses to each other, the null hypothesis H_0 and the alternative hypothesis H_1 [51]. Depending on certain decision rules, the alternative hypothesis is either accepted in favor of the null hypothesis or rejected, maintaining the null hypothesis. For instance, H_1 can typically be that "There is a linear relationship between these two data sets", while the respective H_0 is "There is no linear relationship between the data sets".

The null hypothesis is assumed true and only rejected if the statistical test determines a level of statistical significance for the alternative hypothesis based on the collected data [52]. When analyzing the data statistically, the p-value is determined, and indicate the probability of observing significant values of correlation by chance if the null hypothesis is indeed true. The significance level, α, is used to set a threshold for keeping or rejecting the null hypothesis, where α is the predetermined level of statistical significance. If the p-value is lower than α, H_0 is rejected in favor of H_1, and on the other hand H_0 is maintained if p is higher than α.

The significance level has a conventional range between 0.01 and 0.10, with 0.05 as the standard level for significance [52]. This denotes the probability of committing a Type I error or getting a false positive result, i.e. rejecting the null hypothesis when it is actually true. Resulting in an inferred relationship by the analysis when the reality is no shared behavior. Consequently a p-value below 0.05 confer that more than 95% of the time the observed relationship is from a significantly correlated pattern. A false negative or a Type II error, on the hand, is when the alternative hypothesis is true but is rejected, inferring no association when there actually is a correlation.

In the case of comparing gene expression patterns, the alternative hypothesis is that there is a correlation between two genes, and the hypothesis is tested against the null hypothesis stating no correlation. If the difference is within the region of acceptance, the H_0 is kept and the alternative hypothesis is rejected. On the other hand, if the difference of the correlation is in the rejection region, below the significance level α, the H_1 is accepted based on the observed relationship between the gene pair from the expression profiles.

2.4.5 The Problem of Multiple Comparisons

The problem of multiple comparisons arises when testing several hypotheses simultaneously [53]. The significant level α specified for each test no longer reflect the true chance of a Type I error of multiple comparisons, as the probability increases, often sharply, with the number of hypotheses [54].

One such case is when using microarray data for analyses, such as comparing gene expression data sets between diseased and healthy individuals or treatment/control comparisons. These arrays typically consist of tens of thousands of measured genes. When constructing a differential gene co-expression networks, 20 000 measurements compared pair-wise would result in testing close to $2 \cdot 10^8$ hypothesis simultaneously. With a p-value of 0.05 that is normally considered significant, it would allow almost $1 \cdot 10^7$ false positive results. Using these results for an differential co-expression analysis to for example identify genes (or gene-pair relations) involved in disease, one would include a high number of false results and end up with a substantial concealment of the true relations. The conclusions drawn from such a network quickly become rather useless as "evidence". Therefore, it is important to do proper corrections to account for these effects. Two possible ways of controlling this is the Bonferroni correction and the false discovery rate.

The Bonferroni Correction

The Bonferroni correction method is an approach to lower the threshold α to minimize the number of false positive results when doing multiple comparisons simultaneously [55]. This correction is simply done by dividing the original threshold α by the number of tests being conducted. This is method works best when applied to a small number of tests [53].

When the number of tests are high, the Bonferroni correction can make it difficult to make any significant discoveries at all, even the valuable ones. If for example 1000 test are to be conducted simultaneously, the new critical value of α would be $5 \cdot 10^{-4}$, which is low. This makes it difficult to obtain significant correlations with a p-value of $5 \cdot 10^{-4}$ or less. Making even more simultaneous tests, such as when conducting gene-pair comparisons from microarray data sets, any significant results quickly becomes impossible to obtain.

The False Discovery Rate

Another method to correct for multiple comparisons is using the false discovery rate (FDR), which is the expected portion of false positive among all the discoveries [56]. Giving us the percentage of the obtained positive results that truly are negative. This is preferred in fields like systems biology, where the simultaneous tests often exceed several thousands, and the goal of the comparison is discovery and not a limitation of making any errors [53].

2.5 Gene Expression

Every cell is a complex system in which a large variety of different proteins interact and function [6], such as *Escherichia coli* K-12 that can produce just under 4300 proteins [57]. Each of these proteins carries out specific tasks with high precision. Not all proteins are needed at all times, and the cells encounter different situations that require different proteins. Therefore, the cells continuously monitors its internal and external environment by sensing a variety of signal. This in an effort to respond to the signals by producing appropriate proteins that act upon the internal or external environment. When damaged, for instance, the cell produces repair proteins. When sugar is sensed, the production of proteins that transport the sugar into the cell and utilize it begins [6].

The information on how to make these proteins are stored in the cell's *deoxyribonucleic acid* or genetic material, referred to as DNA [58]. The information needed to produce a given protein is encoded by a specific stretch of DNA called a *gene*. The gene is transcribed by *RNA polymerase* (RNAp) to produce mRNA corresponding to the gene's coding sequence. This is then translated into a protein [6].

The rate of transcription or expression of a gene is controlled by regulatory mechanisms. Every gene has a regulatory region of DNA, called a *promoter*, preceding it, where RNAp binds. The rate of transcription of a gene, number of mRNA produced per unit time, is regulated by the binding affinity between RNAp and the promoter. This affinity can be modulated by transcription factors that affect the rate by binding to specific sites of the promoters of the genes they regulate. When they bind, they affect the rate of RNAp initiating transcription of the gene, by changing the affinity between the promoter sequence and RNAp. Transcription factors that increase the transcription of a gene is known as *activators* and transcription factors that reduce the transcription of a gene is known as a *repressors* [6].

These transcription factors are a tool to represent the internal and external state of the cell. They are usually designed to alternate quickly between their active and inactive state, modulated by specific environmental or internal signals. The transcription factors will, after such a signal, regulate their target genes to activate or repress the transcription of appropriate proteins [6].

2.6 Gene Co-Expression Network

In a gene-co-expression network the genes are represented as nodes and the link between them represent that they have a coordinated gene expression pattern. Their correlation can have the patterns explained above (Section 2.4.1); positive correlation, negative correlation, or no correlation of their expression. This is calculated from the correlation between their gene expression vectors from the data sets. The network can be unweighted, resulting in a link and a *co-expressed* gene pair for every correlation above a given threshold, or it can be weighted, where the thickness of the link reflect the strength of correlation and co-expression.

The interest in finding co-expressed genes is that this indicate which genes are active simultaneously, often in the same biological process [59]. These gene co-expression patterns are often found to coincide with the given phenotype or biological functions [8, 9, 10].

Gene co-expression networks makes it possible to identify network characteristics and reveal structures of a complex network of interaction that would be hard to conceive when only looking at the genes or gene pairs separately.

To generate a gene co-expression network, gene expression data is needed, for instance by performing RNA-seq (Section 2.2.1). Each gene on the array is measured simultaneously and several independent samples from the condition or tissue of interest is needed. The complete data set is arranged in a matrix. Each row containing a gene expression vector for a given gene from all the analyzed samples and each column containing the measurements of all genes from one sample. One sample is thus on a certain column, with each measurements at the same place in the expression vector for each gene vector given in the rows. Two of these data sets are used to perform gene co-expression analysis from two conditions or other states of interest.

2.7 Differential Gene Co-Expression Network

Differential gene co-expression analysis investigate how the pair-wise correlations in gene expression differ between two or more conditions, looking for condition-specific co-expression patterns often linked to dysfunctional regulation [60]. There are different ways to achieve this comparison, broadly divided into two different categories. The first category generally make separate co-expression networks, one for each of the conditions [60, 61]. Here, the genes are connected if their co-expression score is considered significant by the given statistical criteria. The networks are then compared to extract interactions present in only one of the networks or to identify genes with substantial rewiring. The second category on the other hand focuses on scoring each gene pair and establishing if the change in co-expression is significant between the conditions [60]. This can be achieved by generating one differential co-expression network, in which the link represents the co-expression relation between the two conditions [11, 62].

Differential gene co-expression analysis can thus be used to identify co-expression patterns that are specific to the diseased tissue, and point out characteristics of this tissue compared to the healthy control. Further it can be used to identify modular structures related to the disease phenotype. Disease modules pointing to phenotype of interest can guide the future experimental work in the direction of uncovering disease mechanisms, predicting disease genes and advances in drug development [41].

2.8 The CSD Method

The CSD method provides a systematic framework for analyzing differential gene co-expression networks. It identifies and incorporates three different types of differential co-expression changes for gene pairs between two conditions, e.g. disease and healthy control. The generated CSD network consist of nodes, representing genes, and links, representing a change or conservation of the co-expression pattern between two genes across the two conditions.

Firstly, the pair-wise gene co-expression scores $\rho_{i,j}$ is calculated separately for each condition, by the Spearman correlation coefficient for genes *i* and *j* over all the *N* gene expression data points in the given tissue *k*. This measure holds quantitative information of the similarity for a the given gene pair *i* and *j* in condition *k*. In the given condition each co-expression relationship can either show a strong correlation (positively or negatively, with $\rho_{i,j}$ close or equal to 1 or -1, respectively) or a weak/no correlation ($\rho_{i,j}$ close or equal to 0).

This is used to quantify the difference between two conditions. The CSD method incorporate the co-expression scores between two conditions to identify meaningful categories of differential expression pattern between the conditions. In the CSD method a co-expression relationship from one tissue may either be similarly co-expressed (C), co-expressed but with the opposite sign (D), or not have any significant co-expression (S) in the other tissue, revealing three co-expression relationships between the two tissues:

- A *conserved* (C) link represents a significant co-expression relationship between a gene pair that is similar in both conditions, i.e. it has not changed. The correlation is strong in both conditions, and has the same sign.
- A *specific* (S) link represent a significant co-expression relationship between the gene-pair in one condition and no or a weak correlation in the other conditions. The correlation is strong of any sign in one condition, but not in the other.
- A *differentiated* (D) link represent a significant co-expression relationship between a gene pair in both conditions, but the sign changes between the conditions. The correlation is strong in both, but with opposite signs.

The three different co-expression relationships identified by the CSD method are visualized schematically in Figure 2.4. The colored regions represent the areas corresponding to the co-expression relationships described above: two blue areas represent C relationships, four green areas represent S relationships and two red areas represent D relationships. These relationships become the link attribute in the resulting CSD network. The white area represent combinations of correlation coefficients ρ_1 and ρ_2 for gene pairs in condition 1 and 2 that will not result in a link in the CSD network. This involves situations were the correlation of one or both pairs in the conditions are neither particularly strong or particularly weak, and where both correlations are weak.

The C, S and D relationship for the gene pair *i* and *j* in the two conditions is determined by calculating their gene relationship scores, C_{ij},S_{ij} and D_{ij} as given in Equations 2.9, 2.10 and 2.11, respectively:

$$C_{ij} = \frac{|\rho_{ij,1} + \rho_{ij,2}|}{\sqrt{\sigma^2_{ij,1} + \sigma^2_{ij,2}}} \tag{2.9}$$

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

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

The scores quantify to what extent the co-expressions for gene pair *i* and *j* are conserved, specific or differentiated respectively. Note that with the absolute value making each numerator positive, and the denominator being a positive number with no limitations of how close to 0 it is, the scores can take any value from 0 to infinity. As the C, S and D score follow different distributions, the values are not directly comparable. To integrate the scores in a comparable way in the same network, the scores have to be combined with suitable threshold values so that each score corresponds to an *importance level*, *p*. Further explained in Section 2.8.2.

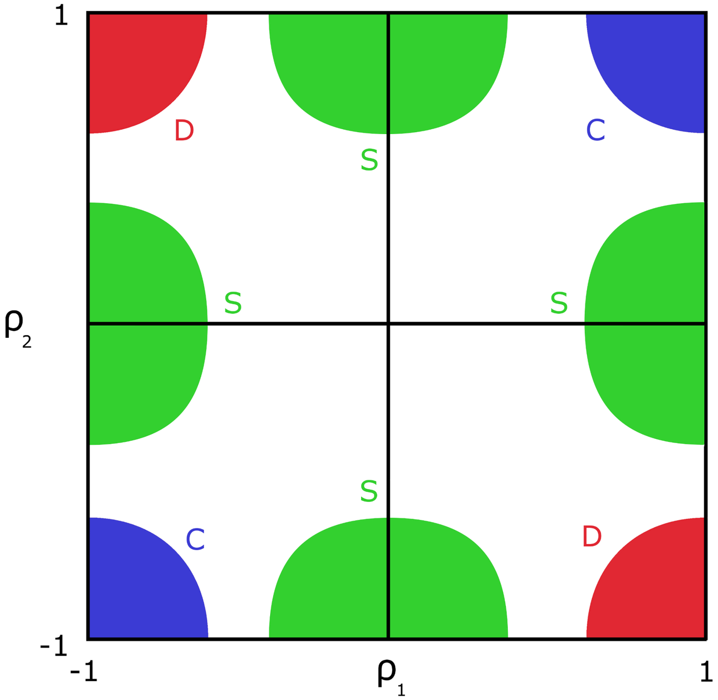


Figure 2.4: General representation of regions corresponding to correlation coefficients from two conditions that results in the three types of differential co-expression relationship, C, S and D. The variables ρ_1 and ρ_2 denotes the Spearman correlations of a given gene pair in condition 1 and condition 2 respectively. The colored region correspond to ares that are included as a link in the network with the color assigning the relationship type: blue is conserved, green is specific and red differentiated. The letter next to each colored region also indicate the co-expression relationship. The white areas are not included in the network. Source: [11]

2.8.1 Variance Estimation in Each Condition

As a way of accounting for potential confounding factors, that may change the correlation within a given subpopulation, the extent of variability of $\rho_{i,j}$ within a condition need to be determined. This is done by including the variable $\sigma^2_{ij,k}$ in the equations calculating the relationship scores (Equation 2.9, 2.10, 2.11). This is an estimate of the variability in Spearman correlation coefficient for each gene pair *i* and *j* in condition *k*. As mentioned in Section 2.4.3, high variability in correlation reflect groups of samples that are differently correlated than other samples, which can be caused by an unknown factor only affecting this subgroup from which the samples are taken. If not corrected for, this could result in a false impression of relevance of the gene pair in the studied condition, when the reality is that it is associated to an unknown factor not accounted for. A high $\sigma^2_{ij,k}$ consequently reduce the value of each the co-expression scores.

This measure of internal variance in the co-expression for a gene pair in each condition is calculated from the set of Spearman correlation coefficients in each independent sub-sample, as the standard error of the mean. The sub-samples are selected by the following algorithm:

1. The complete set of *N* data points per gene are ordered and sequentially numbered.
2. The set is divided into non-overlapping sub-samples of size *n*, e.g. *N* = 69 and chosen sub-sample size *n* = 7 that initially creates 9 sub-samples.
3. Initiating sub-sampling with the first data point (*N* = 1) as the initiating data point $n*$, the current sub-sample is constructed by sequentially iterating through the data points. A data point is added to the current sub-sample if it has not co-occurred with any of the points already in the sub-sample.
4. When the current sub-sample reaches the chosen sub-sample size *n*, a new sub-sample is initiated with initiating data point $n*$ and then repeating step 3.
5. When a sub-sample of size *n* can no longer be drawn using $n*$ as the initiating data point, a new initiating data point is chosen by increasing $n*$ with an increment of one $n*$ ($n* = n* + 1$) and step 3 is repeated with this new initiating data point.
6. The approach is completed when $n* = N$ and no more valid sub-samples of size *n* can be drawn.

In order to increase the chance of matching these sub-samples with a confounding factor, the algorithm ensures that the highest possible number of independent sub-samples of a fixed size *n* is drawn from the complete set. The sub-sample size should therefore be small in order to detect confounding factors as well as possible, while also allowing reasonable calculations of correlation coefficients. Voigt et al. found that a sub-sample size of $n = 7$ to be the minimum requirement, and that the data set should have at least *N* $= n^2$ data points per gene for the estimation of $\sigma^2_{ij,k}$.

Independence of the sub-samples is ensured by the condition that two gene pairs can only co-occur once in a sub-sample. The reasoning for using independent sub-samples is that calculating correlations over the same data points several times can possibly result in an underestimation of the real variability within the data set, ultimately masking potential confounding factors.

2.8.2 Thresholds for the C$_{ij}$, S$_{ij}$ and D$_{ij}$ scores

To map the three scores to a common scale, three threshold values, k_p^C, k_p^S and k_p^D, are determined so that each of them corresponds to an importance level, *p*. This is not the same as the *p*-value in a hypothesis testing situation. Instead the importance level is determined by the distribution of the scores, and the probability of obtaining a given value from these distributions. A collection of values from each score is discarded if they are below this given threshold. This cut-off values should be computed so that none of the areas overlap in respect to each other, and adjusting the importance level corresponds with increasing or decreasing the colored areas in Figure 2.4. Adjustment can be made to ensure a network size and link density that is suitable for further analysis.

The thresholds are calculated in the following way, with the C score and accompanying values as an example: The C score is calculated for all *M* gene pairs from the total set of *N* genes. From the *M* different C_{ij} scores, a sample s_i is drawn *m* times, each with a sample size *L*<<*M*. The threshold k_p^C is determined as the average of the maximum values per sample, given in Equation 2.12. The importance level is set as *p* = 1/*L*.

$$k_p^C = \frac{1}{m}\sum_{i=1}^{m} max_{s_i} C \tag{2.12}$$

2.8.3 Node Homogeneity

The final network generated by the CSD method can have nodes connected to their neighbors by a link-type of either C, S or D. The given node *i* in the network can be characterized by the distribution of these link types, termed *node homogeneity*, H_i, and is given by:

$$H_i = \sum_{j \in \{C,S,D\}} \left(\frac{k_{j,i}}{k_i}\right)^2 \tag{2.13}$$

where $k_{C,i}$, $k_{S,i}$ and $k_{D,i}$ represent the number of C, S and D-type interactions, respectively, of node *i*, and k_i is the degree of node *i*. The node homogeneity quantifies if a node's co-expression relationships predominantly are of one type, or close to an equal amount of every type (C, S and D).

Chapter 3

Materials and Methods

3.1 Data set collection

The gene expression data for healthy tissue used in the analyses of this thesis was downloaded from the Genotype-Tissue Expression (GTEx) project [63] in fully processed and filtered single-tissue gene expression matrices in .bed-format. The data set contained 396 data points per gene. The gene expression data for different the subtypes of breast cancer, and the histologically normal tissue adjacent to tumor was retrieved from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) study [64], with permission from The METABRIC Data Access Committee. The data sets are deposited at `http://www.ebi.ac.uk/ega/`, under the accession number EGAC00001000484. The samples from METABRIC were transcriptionally profiled on the Illumina HT-12 v3 platform and log2-normalized, as described in [64]. The data sets for breast cancer was divided between *discovery* (997 samples) and *validation* (995 samples), with 1992 breast cancer samples in total. The tumor-adjacent tissue data set, *normals*, contained 144 samples.

3.2 Data Integration

For the differential co-expression analysis the compared data sets need to contain corresponding identifiers for each gene, arranged in the same subsequent manner in each data set. The GTEx data set contained Ensembl gene identifiers, while the METABRIC data sets had Illumina (ILMN) HT12 v3 probe identifiers. In order to run the analysis these must match. To identify the ILMN ID's included in the GTEx data set, Ensembl Biomart (`https://www.ensembl.org/biomart/martview`) were used, uploading the stable Ensembl gene ID's and extracting the corresponding ILMN ID's, their Ensembl ID's and their gene symbol. Subsequently, the data sets were reformatted to contain the extracted ILMN probes ID's. The GTEx was reformatted to contain duplicated entries matching several ILMN probe ID's and the METABRIC data sets were filtered to only contain the ILMN ID's available in the GTEx data set. The resulting data sets containing 28361 unique ILMN probe ID's, some referring to the same gene symbol or Ensembl ID. Ideally, the gene name would be the identifier of each measurement, but as the method require unique ID's and some of the ILMN ID's refer to the same gene, these were rather included at a later stage of the analyses.

The data sets of breast cancer from the METABRIC study (discovery and validation) were merged to one set and then divided into the five intrinsic sub-types according to the PAM50 subtype annotations provided in the supplementary information [64]. The resulting data sets had the following composition: 718 samples with the subtype Luminal A, 489 samples with the subtype Luminal B, 240 samples with the subtype HER2-enriched, 330 samples with the subtype Basal-like, 198 samples with the subtype Normal-like.

3.3 Differential Co-Expression Analysis Workflow

Originally the CSD framework developed by Voigt [11] was intended for the differential co-expression analysis of the data sets, hereby referred to by CSD_O. This code is written in C++ and is available from `https://github.com/andre-voigt/CSD`. However, due to time constraint, a faster implementation written in R, hereby referred to as CSD_R, was used. This approach uses bootstrap re-sampling and integrate already existing packages in R for swift calculations. This code is available from `https://github.com/AlmaasLab/csdR`. To use this code, a separate R script was written, available in Appendix A.3, sourcing the function of the CSD calculations. The separate script was written to import and transpose the data sets to the appropriate format, to implement a selection of the resulting C, S and D scores after the CSD calculations, and to generate histograms for the distribution of the scores.

Before being a part of this thesis the R code had not been used analytically, and in cooperation with Jakob Peder Pettersen, the writer of the CSD_R code, appropriate corrections and changes were made consecutively to being discovered when running the analysis. This included equation error in variance estimation, correlation used for variance estimations, and implementing the Welfords algorithm for variance calculations in order to surpass the round-off error causing the estimate of the variance, and consequently the standard deviance, to be set to zero, resulting in Inf-values in the resulting C, S and D scores. When these corrections had been made, the CSD_R calculations was performed by running the code with a given number of bootstrap iterations (B). Here, the pair-wise Spearman correlation and the variance in each bootstrap selection is calculated. Ultimately using the mean of these calculated statistics to calculate the link scores of the gene pairs, as given in Equations 2.9, 2.10, 2.11. After obtaining satisfactory stable result, as described in the following paragraphs, the top 1000 link scores for each link type was filtered out to generate a network suitable for analysis. These calculations were performed on the following combinations of METABRIC and GTEx data sets: histologically normal cancer adjacent (HNCA) and healthy controls (HC) (HNCA:HC), Luminal A and HC (LumA:HC), Luminal B and HC (LumB:HC), Basal-like and HC (BL:HC), HER2-enriched and HC (HER2:HC), Normal-like and HC (NL:HC), resulting in six different CSD-networks.

When performing the calculations using bootstrap, it is important to compare the results of different Bs to ensure stable results. This was done by by running several analyses on the same data sets with different numbers of B, comparing the correlation and variance of the resulting links until the observed values was stable. The seed number was set to the number of iterations performed to ensure reproducible results, for instance when running the CSD_R with 40 iterations, the seed was set to 40, similarly it was set to 100 when running 100 iterations. Using different seed numbers for different B's ensure that the results aren't masked as more similar or stable then they actually are. The same seed number would make re-samples of the overlapping number of iterations identical (i.e. 20 of the re-samples would be identical when running 20 and 40 iterations).

Running the bootstrap with different B's for the BL:HC CSD calculations it was clear that the correlation and the variance was stable between 20 and 40 samples of identical gene pairs. However, this revealed that the number of identical gene pairs selected between the two B numbers (20/40) was very low (4.1%). Consequently, higher number of Bs were tested to gain more comparable results for the stability and identical gene pairs with S-scores, comparing Bs of 200/400, 1000/2000, 2000/4000, 4000/8000, 8000/16000, available in Table 4.1. With the requirement of 66% of matching gene pairs and stability of the correlation and variance of these values, Bs of 2000 and 4000 was selected for the analysis of BL:HC. This was also used as the starting point for the other CSD calculations. For each computation, the selection of links was reviewed for stability in correlation and variance, and the number of matching links between the two was identified, available in Appendix A.4. B was increased if the results were not stable or the percentage of matching gene pairs were less than 66%.

To further look into the few matches of the S links in the BL:HC 20/40 comparison, histograms were generated for the complete link score distribution for each of the link types, at 20 and 4000 Bs. This was also done for the following 4000 B calculations for the S link type, and all histograms are available in Appendix A.5.

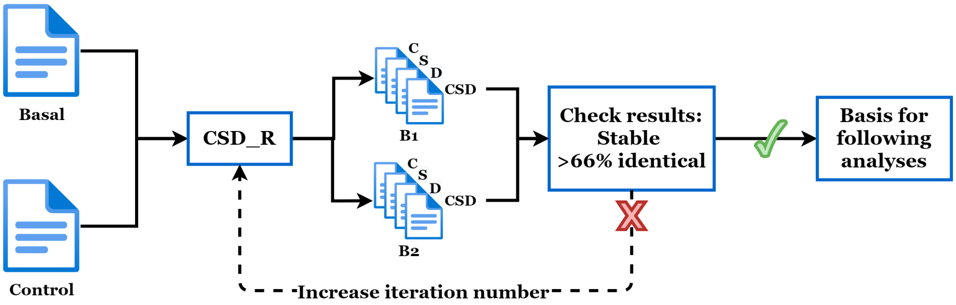


Figure 3.1: Illustration of the work-flow to identify the basis of iteration numbers (B) for stable correlation and variance, in addition to reach a minimum of two thirds identical gene pairs between the comparisons (B1, B2)

After running the calculations on all the data set combinations and inspecting them, they were imported to Cytoscape version 3.7.2. This was followed by adding the link type of interactions and the gene names, which were used to color the edges according to their link type and labeling the nodes to simplify further analyses of the networks.

3.3.1 Evaluating CSD_R network validity

The original plan was to compare the generated networks with CoDiNA networks [65], using wTO-package to generate the input [42], to assess if selecting the top 1000 links is reasonable. Several attempts were made to generate wTO-networks with a reasonable amount of significant links but because of computational running time this was not achieved. Instead, the original CSD framework by Voigt (`https://github.com/andre-voigt/CSD`) was employed using 50 of the samples from each intrinsic subtype and control data set, a considerably reduced workload, to assess the arbitrary choice of keeping the 1000 highest link scores of each interaction type.

Firstly the correlation and variance for each of the data sets is computed as the Spearman correlation by using "FindCorrAndVar" and specifying the file name, the number of genes and the number of data points at the start of the code. The output file can also be named as desired. The number of samples per sub-sample used to calculate the variance could also be altered, and was set to 8. The code was then compiled and run for each of the seven data sets. The sub-sampling algorithm used in these calculations is described in Section 2.8.1 and each sub-sample should have at least 7 samples to estimate the Spearman correlation with a three-digit accuracy.

The output from this first calculation generated the co-expression for each condition, and the files contained the gene pairs and their correlation and variance. The next step was using the python script "FindCSD" to compare correlation of gene pairs between two conditions and calculate the C, S and D scores. The file names of the input was set to the appropriate file name of data sets to be compared, and the output names was set according to the compared conditions. Running the python code produced four output files. One file contain the correlations and variance under both conditions, as well as all C S, and D scores for all the gene pairs. The three other files contain one of the link types, and these three are used to generate the network in the next step.

Using the three files from the previous step as input, the network is generated by using the python script "CreateNetwork". The links included in this network are those with a C, S or D score above the computed threshold, corresponding to an importance level, *p*, for each link type $k_p^{C,S,D}$. In this code the importance level is set by the parameter *selSize* which is equal to 1/(desired p-value). The script generates four networks, one for each interaction type and one collected network with all three link types. Different importance levels were used in order to assess the node count and linkage between different CSD_O and the corresponding CSD_R. Resulting networks were imported to Cytoscape 3.7.2 for visualization and analysis.

3.4 Gene Ontology Enrichment Analysis

To assess enrichment of biological processes of the differentially co-expressed genes for each of the networks, gene ontology (GO) biological process enrichment analysis was performed for the complete set of genes in each network. The Ensembl ID was used to perform the GO enrichment analysis using the GO Enrichment powered by PANTHER [66] (available from: `http://geneontology.org/`) and DAVID [67] (available from: `https://david.ncifcrf.gov/`).

The result is a list of biological processes with significant *enrichment* for the genes given as input, assembling the most pertinent terms of the gene set [68, 69]. A biological process term is enriched in the set of input genes if the genes connected to the term is under- or over-expressed compared to what is expected by chance. In order to correct for multiple comparisons, only results with an FDR lower than 0.05 is included. The result also list the number of genes in the given process, a fold enrichment representing the extent of the over- or under-representation in the input gene list.

3.5 Investigation of Network Modules

To investigate modules for functionality and potential disease associations network community detection was performed on all of the networks constructed by CSD_R. This was done by using the Python *community* package which uses the Louvain-algorithm to identify communities [70]. Within the package the "best partition" function was used, and modules were written out to a file and saved as an attribute to the nodes of the networks. Subsequently the nodes affiliated to a module with six or more nodes were colored according to their module affiliation, and are available in Figure 4.3 and Appendix A.8. The detection of network communities aims to identify functional modules and potentially identify disease modules. The bigger modules were further investigated for the presence of disease genes and involvement in biological processes.

3.5.1 Identification of Disease Genes and Potential Disease Modules

Disease genes associated with breast cancer in each network were identified by DAVID, using the entire list of Ensembl IDs from each network, mapping genes to the Gene Association Database (GAD). The genes identified related to breast cancer were annotated in the network and their presence in the modules was investigated.

The larger modules in the network of the network with six or more nodes were further investigated in the aim of identifying functional and potentially disease modules. A GO enrichment analysis was performed for each of the modules using PANTHER and only including significantly enriched biological processes with a FDR < 0.05 with over a ten fold enrichment. Modules showing no significant results with the given FDR were not included in further analysis.

Chapter 4

Results and Analysis

The results in this chapter is divided into three parts according to different part of the study. The first section present and take a closer look at the aspects of using the CSD_R calculations for construction of a differential gene co-expression network. Section 4.2 present the result from the CSD analysis of histologically normal cancer-adjacent tissue and healthy controls and the further analysis of this network. The last section contains the results of the CSD networks constructed from the breast cancer intrinsic subtypes, and the following analysis of these networks.

4.1 Using CSD_R

4.1.1 Selection of iteration number

When using bootstrap iterations, it is important to look at the stability of the calculations to ensure reliable results that do not change between each analysis. This is done by comparing the correlation and variance values for each gene pair between calculations with different bootstrap iterations. Looking at the calculations for gene pairs done with 20 and 40 iterations for the BL:HC network it was apparent that the correlation and variance was stable from the start for identical gene pairs. However, the comparison revealed that there were few identical gene pairs with the S link type between the two calculations. Only 41 gene pairs were present in both calculations, while considerably more matched when comparing the gene pairs of the C and D link type, see Table 4.1. This resulted in setting the requirement that 66% of the gene pairs between comparisons had to match for the bootstrap iteration number to provide adequate stability in the results.

This requirement called for calculations with a higher iteration number in order to increase the matching percentage of gene pairs of both the S and D link type. The gene pairs of the D link type met this requirement when comparing results of the calculations with 200 and 400 bootstrap iterations, while the gene pairs of the S link type required 2000 bootstrap iterations. Looking at the correlation and variance between the results of calculations with 2000 and 4000 iterations these were stable. These iteration numbers were set as the baseline for running CSD_R to generate the other networks, and the stability evaluation and the number of matching gene pairs between results of different bootstrap iteration numbers are available in Appendix A.4.

For the remaining networks, the collection of gene pairs with C and D link types met the requirement of 66% matching gene pairs when comparing calculations of 2000 and 4000 bootstrap iterations. This was not the case for all collections of gene pairs with the S link type, and an increased iteration number was required for the LumA:HC and the LumB:HC network. With an iteration number of 16000 and 8000 for the LumA:HC and the LumB:HC networks respectively, the requirement was met for the S-linked gene pairs as well. Comparing the stability of each of the network calculations that met the requirement of matching genes of each link type, showed stable values for correlation and variance.

Table 4.1: Overview of the gene pairs in the top 1000 selected conserved (C), specific (S), and differentiated (D) link score from the Basal-like vs healthy control calculations that, when comparing different bootstrap iteration numbers (B), are identical in both selections (Matches), and the percentage (%). The percentage of S link scores that are identical in both selections is made bold.

B	**Link type**	**Matches**	%
20/40	C	891	89.1
	S	41	**4.1**
	D	552	55.2
200/400	C	967	96.7
	S	283	**28.3**
	D	869	86.9
1000/2000	C	982	98.2
	S	524	**52.4**
	D	945	94.5
2000/4000	C	988	98.8
	S	677	**67.7**
	D	955	95.5
4000/8000	C	994	99.4
	S	746	**74.6**
	D	971	97.1
8000/16000	C	994	99.4
	S	827	**82.7**
	D	985	98.5

To take a closer look at the scores, distributions of all scores for each link type of the CSD_R calculations were plotted for the BL:HC with 20 and 4000 iterations, available in Appendix A.5. Looking at the S score distribution of the BL:HC calculations there is a small interval of the link scores, with a max value of 45 in the calculations of 20 iterations. Looking at the calculations with different iteration numbers, the max value of the results drops to 28.2 for 40 iterations, and in the calculations with 1000 iterations the max value of the S link score is below 20. The few identical gene pairs observed when comparing results of calculations with 20 and 40 iterations may be a result of this small score interval. Between 20 and 40 iterations there are only 41 identical gene pairs of the S link type, and the max value of the link scores drops from 45.5 to 28.2, indicating varied scores for the links. Variation in scores, accompanied with the small score interval, makes it plausible that the score values change enough to change the gene pairs with the top 1000 links scores. With a higher number of bootstrap iteration the matched gene pairs gradually increase, indicating more stable score values. This point to that small score intervals may require a higher iteration number, in order for the variability in the link scores to minimize and produce the same gene pairs between different calculations.

For the C and D scores the score intervals are larger. The C score interval is considerably larger with a maximum value of 4781 at 20 iterations. For all iteration numbers this stays above 4200 and the number of identical gene pairs are close to 900 at the comparison between 20 and 40 iterations. The D link score interval is closer to the S link score interval with a maximum value of 66.6 for calculations with 20 iterations, but drops considerably less, staying above 56 for all iteration numbers. The number of matching D-linked gene pairs is also below the set threshold for the 20/40 comparison with 55.2% matching. This emphasizes that a small link score interval when using CSD_R can result in varied results in the included gene pairs with fewer iterations and that a higher iteration number is needed.

The other calculations were made with 2000 iterations as the starting point and compared with calculations done with 4000 iterations. Three of the network calculations (HNCA:HC, HER2:HC and NL:HC) met the requirements of 66% matching gene pairs with these iteration numbers, while LumA:HC and LumB:HC required a higher iteration number for the S linked gene pairs to meet this requirement, see Table A.3. All calculations had large score intervals for the C score, with the max score between 1065 to 3352, while the S and D score were on the smaller side with the max S scores between 18 to 27.9 and the max D scores between 39.9 to 64.7.

4.1.2 Inclusion of links in CSD_R

CSD_R calculates the CSD link scores using bootstrap re-sampling, but has no selection or filtering of the link scores. The published CSD_O sets a threshold for each of the link scores so that they correspond to an importance level, *p*, see Section 2.8.2. The CSD_R does not include such a threshold and the selection of links was set to the top 1000 links from each link type C, S and D. This inclusion of links is arbitrary, and not determined by an algorithm or threshold for which values to keep. The goal is to construct a network big enough for meaningful analysis, but small enough to be suitable for analysis. To assess this selection of links and the content of each networks, separate networks were also constructed by using the CSD_O with only 50 samples included in the calculations. These were intended for reference and not for gene or module analyses, as they do not represent the complete sample sets.

Given the difference in data input the results is not expected to be identical and no direct comparison can be made. These CSD_O calculations is rather an approach to assess the validity of choosing 1000 links of each interaction type, by taking a look at the link type distribution in the CSD_O calculations. In this assessment CSD_O networks with different importance levels, *p*, were generated and compared with the corresponding CSD_R network. An overview of the number of nodes, edges, and the percentage of each link types in the CSD_O networks with their corresponding importance value is available in Table 4.2. Generally, looking at the networks with around 3000 links, the distribution of links in the link types show that just below 27% of the links are conserved, while about 35% is specific and close to 38% is differentiated.

Table 4.2: The different CSD_O networks generated with a given importance level (p) and the number of nodes and edges (Size) for each network, as well as the percentage for each of the link types: conserved (C), specific (S) and differential (D).

Network	p	Size	C	S	D
HNCA:HC	10^{-5}	4883 nodes and 5457 edges	27.1%	35.3%	37.6%
	$5 \cdot 10^{-6}$	3154 nodes and 2691 edges	26.9%	35.2%	37.9%
BL:HC	$5 \cdot 10^{-6}$	3947 nodes and 2747 edges	27.7%	34.9%	37.5%
LumA:HC	$5 \cdot 10^{-6}$	3928 nodes and 2785 edges	26.4%	38.1%	35.5%
LumB:HC	$5 \cdot 10^{-6}$	3946 nodes and 2752 edges	26.9%	35.5%	37.6%
HER2:HC	$5 \cdot 10^{-6}$	3663 nodes and 2677 edges	26.3%	33.1%	40.6%
NL:HC	$5 \cdot 10^{-6}$	3819 nodes and 2706 edges	25.9%	35.8%	38.3%

4.1.3 Multiple IDs representing each gene

Some of the genes in the network is represented by several nodes, as each unique ILMN ID can be included, pointing to the same gene name and Ensembl ID. The BL:HC networks consist of 50.6% unique IDs (1949 of 3851), with 1 to 14 additional nodes pointing to the same gene name. This was also the case for the other networks with the following percentage of unique IDs 79.9%, 49.3%, 49.0%, 57.7%, and 60.7% for HNCA:HC, LumA:HC, LumB:HC, HER2:HC, and NL:HC respectively.

4.2 Histologically Normal Cancer-Adjacent Breast Tissue

The gene expression data of 28361 ILMN probe ID's from HNCA tissues and healthy controls (HNCA:HC) resulted in a CSD network of 1167 nodes (genes) and 3000 links (gene-pair relations), and is visualized in Figure 4.1, after filtering out the 1000 highest scores of the three link types C, S and D. The majority of the network is made up by the giant component, which consist of 917 nodes (78.6%) and 2857 (95.2%) links. The rest of the nodes are part of components of seven nodes or less.

Most of the links of the giant component are specific and differentiated, and there are visible topological groupings within the component. A specific link in the network represent a strong co-expression relation that is only present in one condition and not in the other: it is condition dependent/specific. The differentiated links represent a strong co-expression relation in both conditions, but with the opposite sign. The conserved links represent co-expression relations which are strong and consistent between the conditions. Following sections will characterize the network and look closer at genes and modules that may represent changes between histologically normal cancer-adjacent tissue samples and those of healthy individuals.

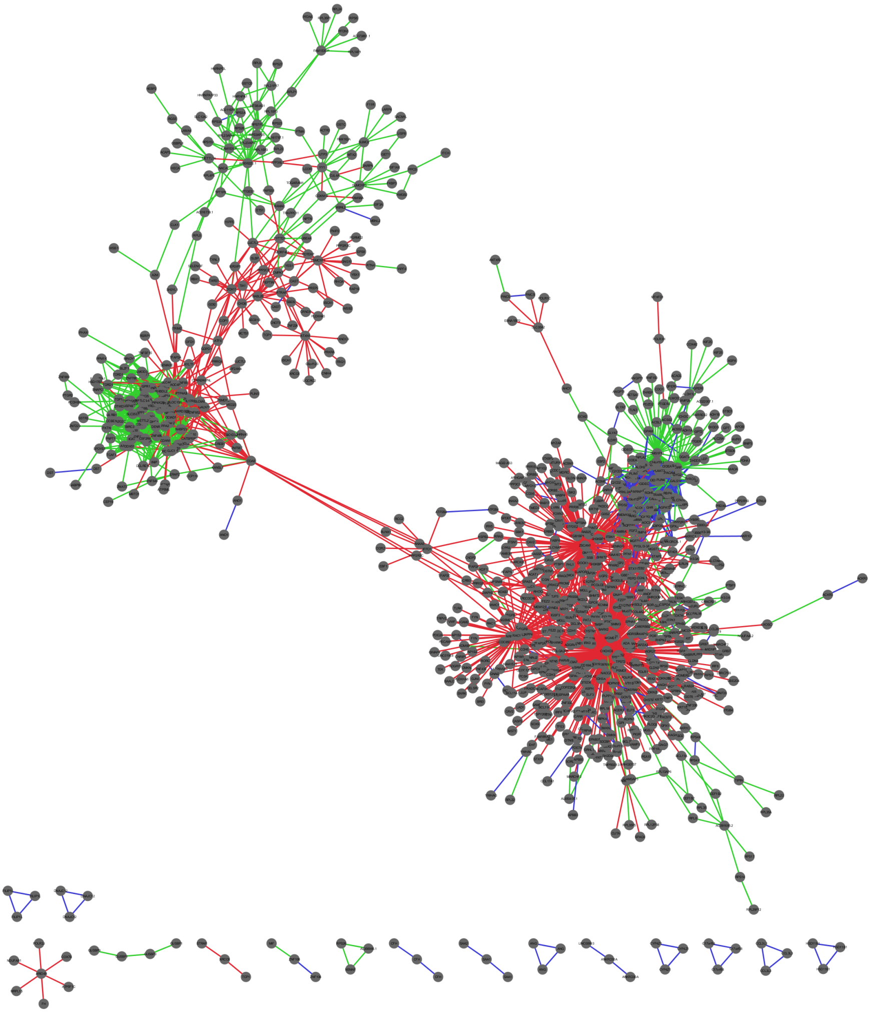


Figure 4.1: Visualization of the HNCA:HC network generated with 4000 bootstrap iterations. Links are colored according to their link-type: conserved links are blue, specific links are green and differentiated links are red in correspondence with Figure 2.4. For visual purposes nodes only connected to another node was excluded from this visualization, excluding 200 nodes and 100 edges.

4.2.1 Degree Distribution

The node degree distribution describes how the network structure is, and the node degree distribution of the HNCA:HC network follow a power-law given by the equation $y = 167.82x^{-1.203}$, and is shown in Figure 4.2 on a logarithmic scale. The good fit of the power-law function indicates that the network is scale free, with a few nodes with a high degree and the majority of the nodes with few neighbors. This is different from the randomly generated networks, in which most nodes are close to the average degree and no hubs. In scale free networks like this one. the hubs have a central role and have important metabolic roles.

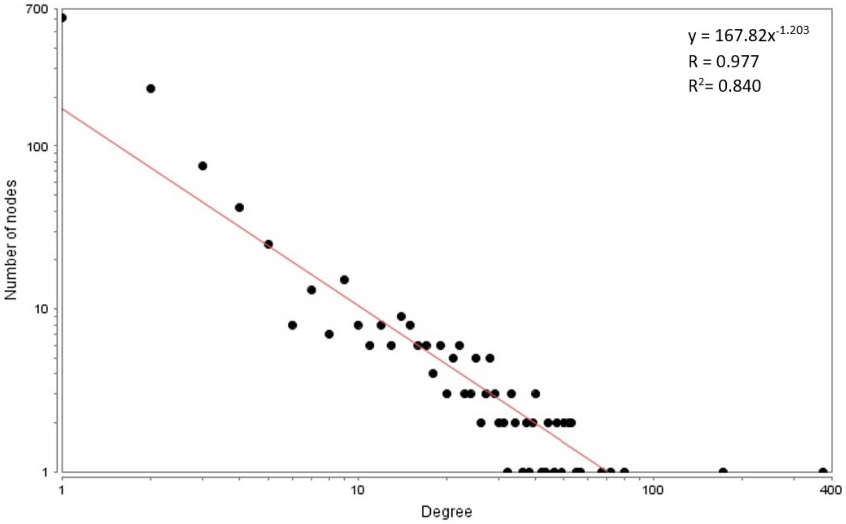


Figure 4.2: The degree distribution plot on a log-log scale of the CSD network of histologically normal cancer-adjacent tissue and healthy controls. The red line representing the approximated power-law fitted function of the data points, with the expression and correlation given in the top right corner.

4.2.2 Hubs and Assortativity

Hubs are defined as highly connected nodes. This is a loose definition without a certain limit set to the degree of nodes to include in the network of interest. A limit set to $k \geq 40$ produced 24 hub genes with degree 40 or higher, listed in Table 4.3 with associated properties like the number of each link type and the homogeneity. Hubs are of biological interest as they represent genes that are co-expressed with many other genes and have functional importance.

To look closer at which link type dominates the hubs, the number of each link type and the node homogeneity H is given for each hub in the table. They correspond to 1.9% of nodes in the network and all of the hubs are dominated by a certain link. Each of the link types C, S and D dominates at least four hubs each. This indicate that there are hubs involved in co-expression patterns that have remained the same, that have been lost or gained, or that have become the opposite.

Table 4.3: The genes in the CSD network of HNCA tissue and healthy controls categorized as hubs, using $k \geq 40$ as the limit. For each hub total degree k, the number of each connection type, k_C, k_S, k_D, the node homogeneity, H, and the dominating link type, H_{dom}, is also given.

Gene	$\boldsymbol{k}$	$\boldsymbol{k_C}$	$\boldsymbol{k_S}$	$\boldsymbol{k_D}$	$\boldsymbol{H}$
CHCHD3	374	0	13	361	0.93
ZSCAN1	172	0	0	172	1
RAC1	80	0	0	80	1
BMS1P10	72	0	72	0	1
ADIRF	67	0	67	0	1
LIPE	57	55	2	0	0.93
GPD1	55	52	2	1	0.9
PLIN1	53	51	2	0	0.93
AQP7	53	51	1	1	0.93
TRARG1	52	50	2	0	0.93
CIDEC	52	50	2	0	0.93
AQP7	50	48	1	1	0.92
CIDEC	50	48	2	0	0.92
MTARC1	49	47	2	0	0.92
AGPAT2	47	43	2	2	0.84
AGPAT2	47	43	2	2	0.84
RBP4	46	44	2	0	0.92
CALB2	44	41	2	1	0.87
GYG2	44	42	2	0	0.91
CFAP74	43	0	43	0	1
ACO1	42	37	2	3	0.78
PFN1	40	0	0	40	1
EID2B	40	0	40	0	1
CEBPA	40	38	2	0	0.91

The differentially dominated hubs of the network are involved in processes of energy production, gene expression, proliferation and apoptosis and motility. *CHCHD3* is protein coding for an inner mitochondrial membrane protein essential for mitochondrial function and a loss of *CHCHD3* expression leads to defects in energy production and cellular metabolism characterized by reduced cellular oxygen consumption and glycolysis rates [71]. *ZSCAN1* is a protein coding gene for a zinc finger and SCAN domain-containing protein that may be involved in gene expression and transcriptional regulation [72]. *RAC1* is also a protein coding gene and encodes a Rho GTPase, which is involved in a multitude of processes such as gene expression, proliferation, apoptosis, inflammation, and regulating the cytoskeleton, and is linked to cancer [73, 74]. It is also suggested as a crucial factor in different types and stages of the inflammatory responses [75].The last hub dominated by differentiated links is only linked linked by this type. It is *PFN1*, which encodes a key regulator for actin polymerisation, that is involved in many processes like motility, signal transduction and gene transcription [76].

The hubs dominated by the S link type is described in this paragraph and represent central interaction that are lost of gained between the healthy samples and HNCA samples. *BMS1P10* is the BMS1 pseudogene 10. *ADIRF* is a protein coding gene for the adipogenesis regulatory protein which plays a role in early adipogenesis and possibly involved in transcription activation [77]. *CFAP74* is dominated by specific links and encodes a protein containing a ASPM-SPD-2-Hydin domain, which is commonly associated with cilia, flagella, centrosomes, and Golgi bodies, and binding microtubulies [78]. *EID2B* is exclusively linked by specific links and is a protein encoding gene. The EID-2B is a transcription repressor and an inhibitor of differentiation [79].

The following hub genes are dominated by conserved links, indicating a gene co-expression pattern that is maintained in both conditions. The common denominator of the hub genes in this paragraph is expression in adipose tissue and involvement in lipid metabolism and energy homeostasis. *LIPE* encodes the protein lipase E, or hormone-sensitive lipase (HLS), that is expressed at high levels in adipocytes (fat cells), which are the primary component of adipose tissue and specialized in storing fat [80]. HLS catalyzes the rate-limiting step in lipolysis, the breakdown of triacylglycerols (TAG), and is therefore critical regulator of energy homeostasis. *GDP1* has a central role in carbohydrate and lipid metabolism as it catalyzes the reversible conversion of dihydroxyacetone phosphate (DHAP) and reduces nicotine adenine dinucleotide (NADH) to glycerol-3-phosphate (G3P) and NAD+, as well as a functioning in the transport of reducing equivalents from the cytosol to the mitochondria (e.g. NADH). [81]. *PLIN1* encodes the perilipin 1 which is involved in adipocyte lipid metabolism and located on the surface of lipid droplets (LD) that store triglycerides and makeup the main energy storage depots in the body [82]. The LDs are important in regulating lipid and glucose metabolism, and perilipins and HLS, mentioned above, is the most important participants in lipolysis. While HLS commence TAG breakdown when stimulated, perilipins on the surface of the larger LD in mature adipocytes mediate interaction between the droplet and HSL when phosphorylated by proten kinase A and stabilizes the LDs and both are important for a optimal lipolysis. *APQ7* encodes aquaporin-7 that is also involved in fat and glucose metabolism in adipose tissue and is responsible for glycerol permeability [83]. *TRARG1* is a positive trafficking regulator of GLUT4 (glucose transporter) in adipose tissue, by promoting translocation to the plasma membrane [84]. *CIDEC* also encode for protein most highly expressed in adipose tissue, with an increased abundance during adipogenesis (formation of adipocytes) [85]. It has been shown to be localized to LDs and promote lipid accumulation.

MTARC1 is also dominated by conserved links and encodes MARC1, an enzyme that, when provided electrons from NADH by Cytb5-R and Cytb5 (electron transport proteins), reduces N-hydroxylated compounds (NHC) and is a component of prodrug conversion and detoxification to avoid accumulation of mutagenic substances [86]. It is also linked to lipid metabolism but the mechanism of MARC1 is uncertain. *AGPAT2* is a protein coding gene and contribute to regulation of lipid metabolism and the enzyme is responsible for synthesizing precursors of phospholipids and TAG and mutations, thought to disrupt adipocyte function, cause congenital generalized lipodystrophy [87]. *RBP4* encodes retinol binding protein 4 that is expressed in liver and adipose tissue and facilitate transfer of small hydrophopic molecules, mainly retinol acid (Vitamin A) [88]. RBP4 can activate pro-inflammatory responses and is correlated with several cancers. *CALB2* encodes calretinin which is normally expressed in retina and sensory pathway neurons, but is also expressed in tissues like adipocytes [89]. It has been detected in 15% of BCs [90], and is involved in processes of intracellular calcium buffering and neoplastic proliferation of the cells expressing it.

Another conserved dominated hub is *GYG2* that encodes a self-glycosylating protein involved in initiation of glycogen biosynthesis [91]. The next hug gene, *ACO1*, is a hub linked by mainly conserved links and is protein coding. It codes for a soluble aconitase which is bifunctional with involvement in iron homeostasis by regulating synthesis of proteins required for uptake, storage and use of iron by the cell, and as the cytoplasmic isoform of aconitase, which converts citrate to isocitrate [92, 93]. The last hub gene *CEBPA* is dominated by conserved links and encodes the transcription factor C/EBPα which is a general inhibitor of cell proliferation and a tumor suppressor, and has been linked to breast cancer [94].

Note that some of the genes or proteins are referred to by synonyms in the cited sources: *ADIRF* is referred to as *C10orf116* with the protein product AFRO, while IRE-BP and IRP1 are synonyms used for the gene product of *ACO1*, and EID-3 is a synonym for EID-2B.

4.2.3 Biological Process Enrichment Analysis

The differentially co-expressed genes of the generated HNCA:HC CSD network showed high enrichment of genes related to several biological processes. The unique genes of the network (933 genes) were mapped to PANTHER/DAVID IDs by their Ensembl gene IDs and 93%/84.7% of these were identified. The general GO biological processes, that generally include a high number of genes, were moderately enriched in the network and regulate a various of processes, listed in Table 4.4. The more specific categories, consisting of fewer genes, were highly enriched and are listed in Table 4.5. The specific biological processes can be categorized into two more general categories: lipid metabolism, and inflammation/immune response.

Most of the enriched biological processes that was specific are related to lipid metabolism, storage and lipolysis. Adipose tissue is one of the main components of the breast, functioning to store excess energy and release it when necessary for the body [95]. Some of the other enriched processes can be linked to inflammation, which is a feature that is increased in tissue within the breast cancer tumor itself and in the surrounding microenvironment.

4.2.4 Network Modules And Disease Genes

In order to identify breast cancer-associated genes that are present in the HNCA:HC network, all Ensembl IDs of the network were submitted to DAVID and 610 of the IDs were mapped to the Gene association Database (GAD), which relates genes to specific diseases. There were no enrichment for breast cancer with a FDR < 0.05. Some diseases were significantly enriched with a FDR below 0.05 and the top three of these when sorting by fold enrichment (FE) were obesity/hypertension (FE = 10.6), blood pressure, arterial hypertension (FE = 7.9) and kidney aging (FE = 3.9).

Table 4.4: The top five general biological processes (BP) enriched in the generated HNCA:HC network and their respective fold enrichment (FD) using the least specific category of biological processes in DAVID. The entries are ordered by their fold enrichment.

General BP	FE
Detoxification	3.7
Locomotion	1.8
Biological adhesion	1.7
Growth	1.4
Immune system response	1.3

Table 4.5: The top ten specific biological processes enriched in the HNCA:HC network with their respective fold enrichment (FE). The processes are ordered according to their FE. *Vitamin C

Biological process	FE
Regulation of adenylate cyclase-activating adrenergic receptor signaling pathway involved in heart process	23.7
Response to L-ascorbic acid*	15.8
Establishment of endothelial blood-brain barrier	15.8
IRES-dependent viral translational initiation	12.93
Protein localization to membrane raft	11.85
Diacylglycerol biosynthetic process	11.85
Negative regulation of granulocyte differentiation	11.85
Positive regulation of macrophage cytokine production	11.85
Regulation of sequestering of triglyceride	11.85
Negative regulation of fatty acid oxidation	9.87

The Louvain algorithm identified 125 communities in the HNCA:HC network, and a visual representation is available in Figure 4.3, where modules with more than five nodes are colored and numbered according to their module and module number. 100 modules consisted of two genes linked together and 13 with three nodes. Further investigation was mainly performed on the 12 modules consisting of more than three genes.

The most highly enriched biological processes of the the modules listed in Table 4.8 were identified by PANTHER and each module is elaborated in following paragraphs. All the included processes are statistical significant, with a FDR < 0.05, and only entries with over a ten fold enrichment are mentioned. Modules that was not enriched for a biological process in this manner is excluded from further analysis.

Module 22: The largest module in the network, making up 53.7% of the giant component, consists of 488 nodes and mainly D links (87.9%). The nodes of this module is colored purple in Figure 4.3. The differential homogeneity indicate that this module represent a structure in the network in which the differential co-expression relation of the gene pairs is of opposite behavior between HC and HNCA samples. From the three enriched diseases mentioned above the module has 21 of the identified disease-associated genes, of which 19 are unique. It is enriched in the processes of establishment of endothelial blood-brain barrier (BBB), protein localization to membrane raft, and positive regulation of phagocytosis, engulfment. The adipose tissue is linked to the BBB by adipokines, which play a role in energy metabolism and immune response, but they can also cross the barrier or modify the physiology by acting on the cells of the BBB [96]. The other two of

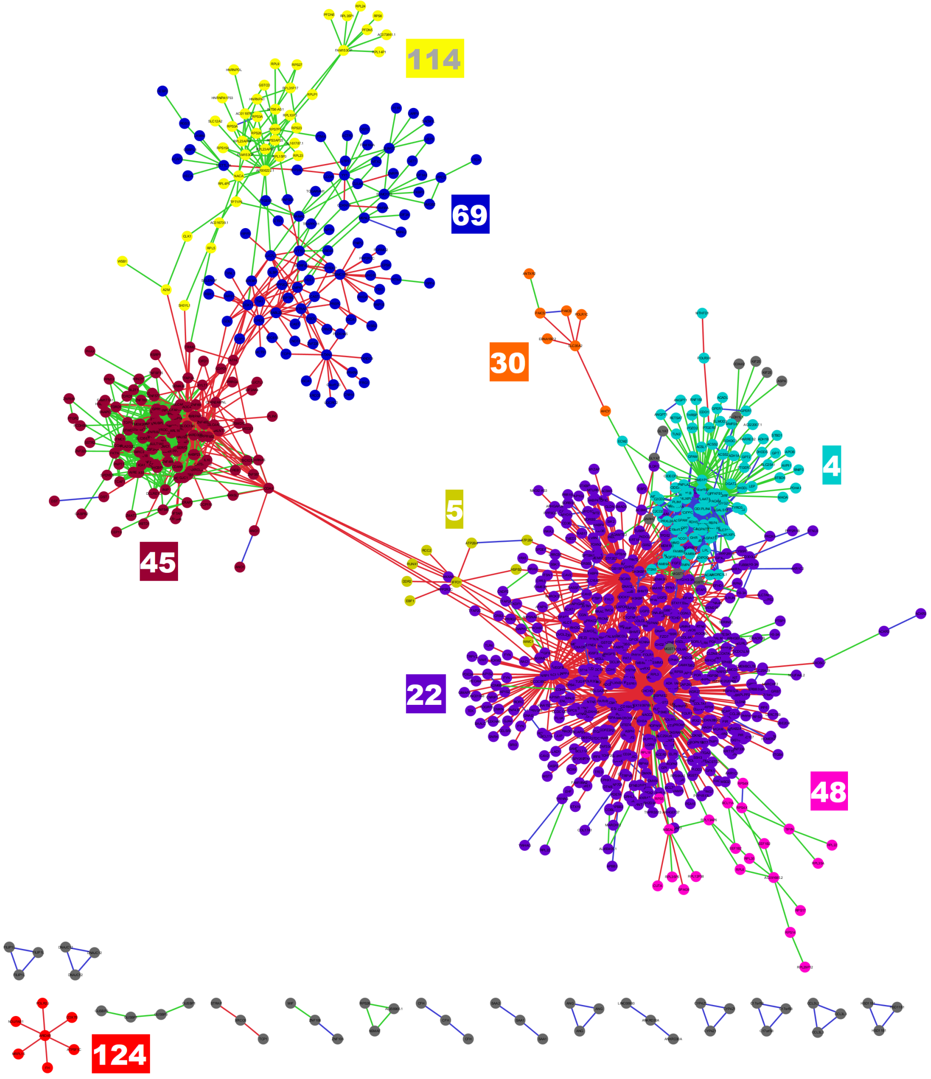


Figure 4.3: Communities detected by the Louvain community algorithm [70] in the CSD network HNCA:HC. Nodes in a community consisting of more than five nodes are colored according to their respective module, and the number denotes the module number. Edges are colored according to their link type, conserved links are blue, differentiated links are red and specific links are green. As in Figure 4.1 genes connected only in pairs are excluded for visual purposes.

the top three processes involve transport of protein to the cell membrane and an increased engulfment of other cells of particles.

Module 114: The yellow module in the top left corner of Figure 4.3 consists of 43 nodes and 80 edges and is a part of the giant component. The module does not contain any of the identified disease-associated genes. It is linked by mainly specific links (96.3%), indicating a differential gene co-expression pattern that is only observed in one of the conditions, and may point to a loss or a gain of relation. Of the 43 genes, 21 were mapped by PANTHER and showed enrichment in negative regulation of ubiquitin protein ligase activity, SRP-dependent cotranslational protein targeting to membrane, and ribosomal small subunit assembly. Ubiquitin is a protein used in many regulatory processes and ubiquitinated proteins are frequently targeted for degradation, but protein ubiquitination may also participate in responses including gene expression, cell cycle, DNA repair and apoptosis [97]. SRP (signaling recognition particle) and its receptor initiate the transfer of nacient secretory protein chain across the ER membrane into the ER lumen [98]. These proteins are initially synthesized on membrane unattached ribosomes, and when SRP bind to the ER signal sequence, they are directed to ribosomes on the ER membrane and into transmembrane channels. The last enriched process is the formation of small ribosomal subunit, which is a part of the translational machinery of the cell.

Module 48: This module is a small part of the giant component, consisting of 22 nodes and 31 edges, visible as the pink nodes at the bottom right in Figure 4.3. It does not contain any of the identified disease-associated genes. Most of the links are of the specific type, and a few links are also conserved and differentiated. The most enriched processes in this module is negative regulation of dendrite extension, ribosomal small subunit assembly and SRP-dependent cotranslational protein targeting to membrane. This module is enriched with two of the same processes as the previous module, and in addition an enrichment for negative regulation of dendrite extension. The dendrite is the brancing of the neuron that receive the synaptic signals from the axon of other neurons [99].

4.3 Breast cancer subtypes

The CSD_R method using gene expression of 28361 ILMN probe ID's from patients with a breast cancer subtype and healthy controls resulted in five networks with 3000 links, 1000 of each link-type C, S and D. General features of the network size and the giant component is summarized in Table 4.6, and BL:HC is visualized in Figure 4.4, while the other networks are available in Appendix A.6.

Table 4.6: Overview of the size of the network and the giant component. Each network consists of 3000 links, 1000 of each link-type.

Network	Size	Giant component
Basal-like	3851 nodes	336 nodes (8.7%) and 916 links (30.5%)
Luminal A	3909 nodes	456 nodes (11.7%) and 1030 links (34.3%)
Luminal B	3859 nodes	418 nodes (10.8%) and 1021 links (34%)
HER2+	2226 nodes	435 nodes (19.5%) and 1953 links (65.1%)
Normal-like	2122 nodes	473 nodes (22.3%) and 1975 links (65.8%)

4.3.1 Degree Distributions

The degree distribution of each of the intrinsic subtype generated networks, BL:HC, LumA:HC, LumB:HC, HER2:HC and NL:HC where found to follow a power-law distribution with a degree exponent γ = 1.952, γ = 1.897, γ = 2.023, γ = 1.507, and γ = 1.504, respectively. This indicate that the networks are far from random with a scale-free topology and central hubs.

4.3.2 Hubs and Assortativity

Using the same threshold for hubs as in Section 4.2.2, the hubs of each of the networks were identified and are listed in Table 4.7 along with their degree, and the number connections of each of the links type. In addition, the homogeneity is listed, to see the extent to which a link type dominates the hubs. Generally the hubs are homogeneously specific or differentiated, with a few conserved or specific links. Some are less homogeneous, displaying a mix of specific and differentiated links.

A hub for four of the networks, *PTP4A2*, encodes a phosphatase involved in the control of cell proliferation and invasion, for which abberant expression is associated with progression and metastasis of multiple cancers [100]. It has oncogenic properties as it down-regulates PTEN expression and thereby activate the P13K-Akt pathway. The P13K pathway is activated by P13K and repressed by PTEN, and the result of an activated pathway is signaling for growth, proliferation, survival, protein synthesis and transcription, as well as inhibition of apoptosis [101]. This hub gene is almost exclusively linked by differentiated links in the four networks and with numerous associations that switch sign between the breast cancer subtype and the healthy control, it is likely to play a role in mediating the disease phenotype.

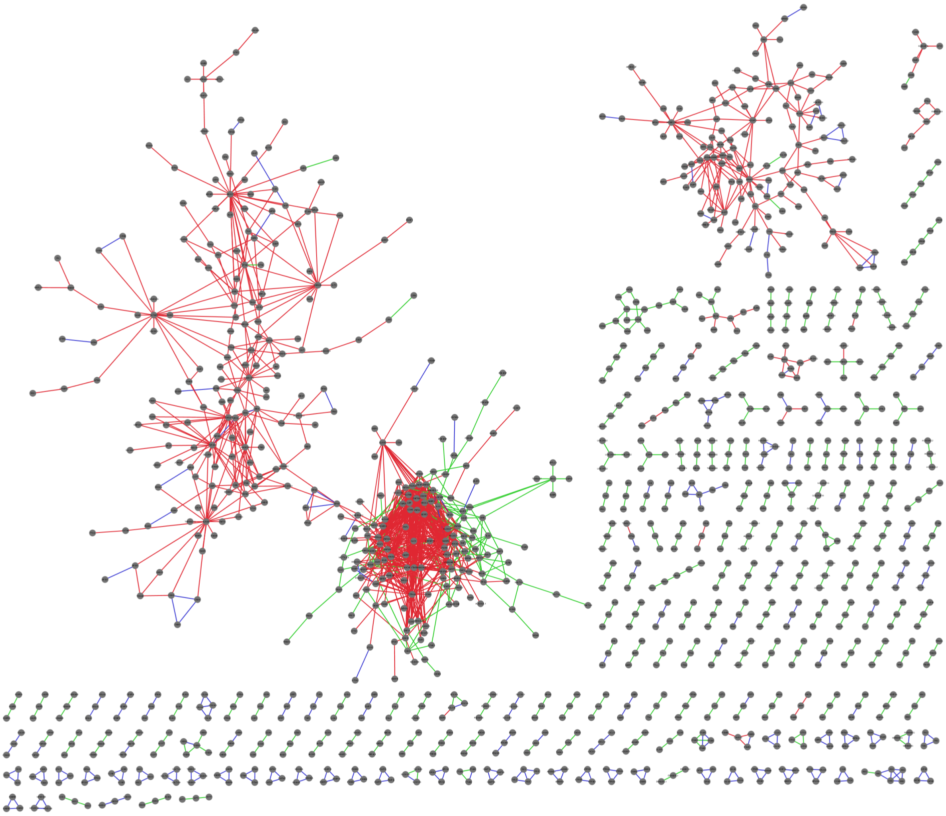


Figure 4.4: Visualization of the BL:HC network generated with 4000 bootstrap iterations. Links are colored according to their link-type: conserved links are blue, specific links are green and differentiated links are red in correspondence with Figure 2.4. For visual purposes nodes only connected to another node was excluded from this visualization, excluding 2670 nodes and 1335 edges.

The *PTP4A2* is the only hub of the BL:HC and the LumA:HC network, while the only hub of the LumB:HC network is *DDX17*, which is a homogeneous hub linked by differentiated links. This makes the gene likely to play a role in the interactions of the disease phenotype. It is protein coding for DEAD box RNA helicase 17 which is known to take part in a range of processes including transcription and RNA processing, as well as deregulated expression in multiple cancers. It has been indicated to have both pro- and anti-proliferation roles in cancer development, likely context-dependent, and is involved in ERα activity and estrogen-dependent growth [102, 103].

Looking at the HER2:HC network, the *PTP4A2* is also a hub of this network and homogeneously differentiated. Another hub dominated by differentiated links is the *TVP23C* hub, encoding the golgi apparatus membrane protein TVP23 homolog C (TVP23C) [104]. It has been shown that higher levels of TVP23C have a more favorable outcome in colorectal cancer patients. The homolouge Tvp23 is thought to be involved in vesicular transport and is associated with the Golgi apparatus membrane [105]. The other differentiated linked hubs, *PSMD12* and *COA8*, also have specific links and are not homogeneous. *PSMD12* is protein coding for the non-ATPase subunit PSMD12 of the 19S regulator of 26S proteasome complex, which is responsible for ATP-dependent degradation of many proteins in the ubiquitin-proteasome system (UPS) of the cell [106]. This system is a biological process in the cell crucial for homeostasis, signaling, and fate determination. *COA8* is protein coding for cytochrome c oxidase assembly factor 8 that is stabilized during oxidative stress, and quickly degraded by UPS otherwise [107]. The function of the protein is to increase and protect cytocrome c oxidase assembly, the last component of the energy producing mitochondrial respiratory chain, from oxidation-induced degradation.

The other hubs of the HER:HC network are homogeneously connected by specific links, indicating a loss or gain of interaction with the genes they are linked to. The first of these hubs are *CFAP74*, which has been described in Section 4.2.2, and contains a domain associated with cilia, flagella, centrosomes, and Golgi bodies, and binding microtubulies. The last hub of this network is *GPR1* which encodes the G-protein coupled receptor 1. G-protein coupled receptors are a large family of transmembrane receptors, which, upon binding of its ligand, modulates intracellular pathways [108]. Most cases involves activation of G-proteins, but it can also occur independently of G-proteins. GRP1's functionality has been shown to be involved in regulation of glucose homeostasis [109]. Additionally it is linked to higher expression in breast cancer and tumor growth [110].

The NL:HC network contains several hubs. *PTP4A* is already discussed above and is mainly linked by differentiated links. *GJC1* encodes connexin 45 (CX45), which is a part of the gap junction. The connexins have highly conserved regions, but differ in their intracellular domains which indicate specific biological properties [111]. The connexins provide direct interaction between adjacent cells and coordination of cellular processes, including growth, and with their different functions, different connexins can induce pro- or anti-tumorigenic effects [112]. *CDKN2AIPNL* is linked by both specific and differentiated links and encodes CDKN2A interacting protein N-terminal like. It is a putative participant in the cell cycle and involved in signal transduction [113]. The next hub, *CFAP74*, is already described in Section 4.2.2, and in this network the hub is homogeneously linked by specific edges. *SLC2A12* encodes the facilitative glucose transporters 12 (GLUT12) that functions as a insulin-dependent glucose transporter, which is a crucial role in glucose utilization and homeostasis [114]. An increase in glucose consumption is a characteristic allowing cancers to grow, and GLUT12 have been detected and implemented in breast cancer [115].

Note that some of the encoded proteins are referred to by alternative names in the cited literature: PRL2 is a synonym for PTP4A2, APOPT1 is an alias for COA8, and SLC2A12 is referred to as GLUT-12.

4.3.3 Biological Process Enrichment Analysis

The differentially co-expressed genes of each CSD network showed high enrichment of genes related to several biological processes. The unique genes of the networks were mapped to PANTHER/DAVID IDs by their Ensembl gene IDs and 94.9%/86.9%, 94.7%/86.7%, 95.2%/87.1%, 95.7%/88.6%, and 95.5%/89.0% genes were mapped from the BL:HC, LumA:HC, LumB:HC, HER2:HC, and NL:HC, respectively. The general GO biological processes, that often include a high number of genes, were moderately enriched in the

Table 4.7: The hub genes identified in each of the CSD networks (BL:HC, LumA:HC, LumB:HC, HER2:HC and NL:HC), their total degree *k*, the number of links of each type, k_C, k_S, k_D, and the node homogeneity, *H*, of each of the hubs.

Network	Hub genes	k	k_C	k_S	k_D	H
Basal-like	*PTP4A2*	44	0	0	44	1
Luminal A	*PTP4A2*	92	0	1	91	0.98
	PTP4A2	67	0	0	67	1
Luminal B	*DDX17*	46	0	0	46	1
HER2+	*PTP4A*	118	0	0	118	1
	PTP4A	74	0	7	67	0.83
	TVP23C	61	2	0	59	0.94
	TVP23C	49	1	0	48	0.96
	CFAP74	47	0	47	0	1
	PSMD12	46	0	10	36	0.66
	GPR1	42	0	42	0	1
	COA8	40	1	11	28	0.57
Normal-like	*PTP4A*	65	0	2	63	0.94
	GJC1	50	0	8	42	0.73
	PTP4A	48	0	0	48	1
	CDKN2AIPNL	43	1	13	29	0.55
	PIP4K2B	41	0	40	1	0.95
	CFAP74	40	0	40	0	1
	SLC2A12	40	0	0	40	1

networks and regulate various processes, like growth, biological adhesion and cell killing. The more specific categories, consisting of fewer genes, were more highly enriched. An overview of general and specific biological processes are available in Appendix A.7 with their corresponding fold enrichment. Most of the specific biological processes could be divided into more general groups of proliferation, apoptosis, immune response, biosynthesis, stress response and motility.

4.3.4 Network Modules And Disease Genes

To identify modules in the breast cancer intrinsic networks, the Louvain algorithm was used. It identified 1555, 1583, 1555, 829, and 638 communities in BL:HC, LumA:HC, LumB:HC, HER2:HC, and NL:HC respectively. A visual representation for each network is available in Appendix A.8, in which modules with six or more nodes are colored and numbered according to their module number. The same modules are listed in Table 4.8. The majority of the modules were made up by a pair of node, or a triplet of nodes, connected to each other, and further analyses are focused on the modules with six or more nodes within each network. An enrichment analysis for each of these modules were conducted in PANTHER to identify highly enriched biological processes. The following paragraphs describe each of these modules with statistical significant enrichment with an FDR < 0.05 and a fold enrichment of at least one ten fold. Elaborations in processes that are not cited is from the definition of the GO term [116, 117]. The identified modules enriched in biological processes may represent functional modules and altered co-expression pattern of modules with cancer-related processes can also be disease modules.

Identification of breast cancer related genes within the networks could point to modules of special interest. To identify these genes each each of the Ensembl gene lists were submitted to DAVID and mapped to GAD. 610, 1337, 1318, 875, and 899 genes were mapped to GAD and part of the disease enrichment analysis. There were only one of the networks with a significant enrichment with a FDR $<$ for breast cancer. This network was the HER2:HC network. For BL:HC and LumA:HC the only disease with an FDR < 0.05 was Type 2 diabetes, while LumB:HC had no significant disease enrichment with an FDR < 0.05. The NL:HC network had several other enriched diseases and the most enriched diseases with an FDR < 0.05 were lymphoma, followed by leukemia and ovarian cancer.

BL:HC modules

Module 42: The second largest component of the network, visible as the orange component in the top right of Figure A.8, consist of 122 nodes and 175 edges. The edges of the module is mainly differentiated, with 18 conserved links and two specific. The only enriched processes with more than a ten fold is the nested processes of B cell activation involved in immune response and lymphocyte activation involved in immune response. These processes are involved in the immune response, with B cells being an antibody-producing cell with the objective of detect and tag foreign antigens (molecules) through the secretion of antibodies that specifically bind foreign antigens, in order for other cells of the immune system to remove it or to activate the complement cascade for elimination by phagocytosis [118, 119].

Module 60: This module consists of the 70 nodes colored green in the middle of the giant component in Figure A.8. The nodes are connected by 133 edges that are mainly differentiated. The most enriched biological processes of this module is positive regulation of establishment of protein localization to telomere, positive regulation of protein and of teleomerase RNA localization to Cajal body and positive regulation of telomere maintenance via telomerase. These processes are important for teleomere maintanance by telomerase, which is thought to be important in cancers [120].

Module 3: This module is a small separate module consisting of six nodes and is linked by 6 conserved edges, positioned close to the middle of Figure A.8. The enrichment analysis show an enrichment in the nested biological processes of T cell differentiation and selection. These processes are involved in the immune response and involves differentiation of progenitor cells in the thymus, followed by a selection of T cells to mature, ensuring that useless or self-reactive T-cells do not mature [121].

LumA:HC modules

Module 25: This module is situated in the middle of the giant component, colored turquoise in Figure A.9, and make up about half of the giant component with 222 genes and 637 edges. The edges are mainly differentiated (95.8%). Enrichment analysis of this module show an over-representation of genes involved in positive regulation of establishment of protein localization to telomere, positive regulation of protein localization to Cajal body, NIK/NF-kappaB signaling, and positive regulation of telomere maintenance. Three of these have already been mentioned in module 60 of the BL:HC network and are involved in teleomere maintanance, protecting the chromosome endings and keeping the strand stable. The other enriched process, NIK/NF-kappaB signaling, leads to processing and release of an active NF-KappaB which is an transcription factor controlling gene expression linked to control of adaptive immunity [122].

Module 1580: The community identified by Louvain is a part of the giant component, visible as the burgundy part at the bottom of Figure A.9. It consists of 53 nodes and 66 edges, of which are mostly differentiated. The modules is only enriched in process of collagen fibril organization, which involves any process that determines the size and arrangement of collagen fibrils within the extracellular matrix.

Module 171: The orange module, shown in the center of Figure A.9, is a part of the giant component and consist of 20 genes and 27 edges that are mainly differentiated. The most enriched processes of this module include caveolae assembly, receptor-mediated endocytosis of virus by host cell, vasculogenesis, and response to estrogen. Caveolae is a plasma membrane raft forming invaginations involved in several cellular processes like cholesterol homeostasis and regulation of signal transduction, but one caveolae Cav-1 is also pointed to as a tumor suppressor [123]. Its inactivation is only associated with ER-positive breast tumors and the inactivation likely results in increased sensitivity to estrogen due to estrogen receptor α up-regulation. In response to estrogen, the ERα translocates to the nucleus and regulates gene expression directly by binding estrogen-response elements (ERE), which in turn promote oncogenic protein expression and inhibition of cell cycle inhibitors, consequently driving breast cancer initiation and proliferation [124].

Vasculogenesis is the process of blood vessel growth, mainly during embryonic development, but also occur from circulating endothelial precursor cells and can contribute to neovascularization in for instance wound healing or developing tumors [125]. The last enriched process is any receptor-mediated endocytosis of a virus by the host cell, and the involved genes are the caveolae genes (*CAV1, CAV2*).

LumB:HC modules

Module 48: This module is visible as the orange section of the giant component in Figure A.10. It is made up by 91 nodes and 171 edges is linked all interaction types, although differentiated links dominates (84.2%). The enriched processes are chondrocyte development and collagen fibril organization. These two processes may be connected in the tumor environment, promoting tumor progression. Mesenchymal stem cells (MSCs) differentiate into cells such as chondrocytes and are known to migrate towards inflammatory sites and also to be incorporated into tumors and interact with them, contributing to tumor growth and progression. Collagen fibrils, on the other hand, are a part of the extracellular matrix and re-organization of the cellular matrix is favorable for invasive tumor cells [126].

Module 818: A smaller part of the giant component, colored navy in Figure A.10, with 15 nodes and 16 edges. All edges are differentiated, except one conserved link. Enrichment analysis show an enrichment in Fc-gamma receptor signaling pathway involved in phagocytosis, leukocyte migration, and adaptive immune response. Leukocyte migration is the movement within or between tissues and organs, and a fundamental immune response that innate and adaptive immune response rely on [127]. These include neutrophils, macrophages and monocytes, which contain Fc-gamma receptors that detect and induce phagocytosis of phatogens [128].

HER2:HC modules

Module number 20: The largest module in the HER2:HC network consisting of 214 nodes and conserved or differentiated link types, with the vast majority being D links (93.9%). The module is a part of the giant component, visible as turquoise in Figure A.11, and make up approximately half of the giant component. The domination of differentiated links indicate that most of the gene co-expression patterns in this module is disturbed between HER2+ breast cancer tissue and HCs. The module also contain six of the breast cancer-associated genes, which are all linked by differentiated links.

GO enrichment of the genes in this module showed an enrichment in positive regulation of establishment of protein localization to telomere, positive regulation of protein localization to Cajal body, and 2-oxoglutarate metabolic process. The first two important processes for telomere maintenance by telomerase, which is thought to be important in cancer progression [120]. The last one is involved in the citric acid cycle involved in energy production and biosynthesis.

Module number 40: This module make up the purple separate component on the right in Figure A.11. It consists of 51 nodes and is linked by every link type. The predominant link type is conserved, followed by specific and lastly differentiated. Noting that the differentiated links are segregated from the rest, and to an extent the same is observed for the specific and conserved links. The module contain two of the breast cancer-associated genes which are linked by specific links. GO enrichment identified an enrichment in many biological processes including T cell selection, positive regulation of cell-cell adhesion mediated by integrin, and regulation of chronic inflammatory response.

Adhesion molecules such as integrins, play a vital role in the immune system [129]. During cancer development they mediate important anti-tumor responses including antigen uptake and activation of tumor-specific T cells and tumor cell killing. However, they can also be used by malignant cells to promote tumor growth by being expressed on the tumor cell - increasing cell proliferation and survival. Promotion of tumor growth is a known feature of the immune system by maintaining chronic inflammation.

NL:HC modules

Module 94: The biggest module of the NL:HC network with biological processes enriched with at least a ten fold and with an FDR < 0.05 consist of 64 nodes and 133 edges. It is visible as the green part in the middle of the giant component in Figure A.12. It is connected by mostly differentiated links, except for eight conserved ones, and the enrichment analysis show over-representation of genes involved positive regulation of establishment of protein localization to telomere, positive regulation of protein localization to Cajal body, and positive regulation of telomerase RNA localization to Cajal body. All of which are processes important for telomere maintanance by telomerase, thought to important in cancers [120].

Module 88: Module 88 is the separate purple component of the top right in Figure A.12, consisting of 22 nodes linked by 25 edges. All but one of these are conserved links, indicating a conserved interaction between the genes of this module in both conditions. The genes are enriched in many processes and the top five are: thymic T cell selection, regulation of type 2 immune response, positive regulation of CD4-positive, alpha-beta T cell differentiation, and positive regulation of interleukin-4 production. These processes are interconnected. The type 2 immunity induce resistance to parasitic infection and activation lead to differentiation of CD4 T cells to type 2 T helper cells. These secrete specific cytokines like IL-4 and induce development and proliferation of other cells that take part in type 2 immunity [130].

Module 270: This module is a separate component made up by 21 nodes and 38 edges that are mainly of the conserved type, and visible as the orange module in Figure A.12. Enrichment analysis using PANTHER identified enrichment in T cell differentiation, positive regulation of T cell differentiation and activation, and the adaptive immune response. The adaptive immune response consists of both T and B cells that both express antigen receptors with discrete antigen specificity, capable of recognizing a potential pathogens [131]. Binding of the given antigen can differentiate T cells into T effector cells that either help the innate or adaptive immune response or that get rid of the virus-infected cell.

Module 96: This module consist of 12 genes, and represent a conserved component of the network linked by 30 conserved edges, visible as the cyan colored component in Figure A.12. The module is enriched in one biological process which is the regulation of cold-induced thermogenesis. This is a process to generate heat to maintain a stable core temperature in response to cold temperatures by increasing metabolism [132].

Module 58: This module represent a conserved of 9 nodes and 12 edges, colored red in Figure A.12. The module is enriched in the processes of synapse pruning, microglial cell activation, regulation of complement activation, and innate immune response. All of these processes can be linked to central nervous system homeostasis. Microglia is the resident immune cell of the brain and may eliminate synapses in response to inflammatory stimulus [133]. The complement cascade is a part of the innate immune system, and likely mediate synaptic remodeling by tagging synapses for destruction.

Module 163: This module is represented by the bright green module in Figure A.12, and is made up by 9 nodes and 9 edges. All edges, except one, is conserved, and the enrichment analysis showed an enrichment in SRP-dependent cotranslational protein targeting to membrane, viral transcription, and nuclear-transcribed mRNA catabolic process. The first process is the SRP mediated targeting of nacient secretory protein chains to the ER-associated ribosomes and into transmembrane channels [98], as reported in module 114 of the HNCA:HC network. The second process is when a viral genome, or a part of the viral genome, is transcribed in the host cell, while the last process is the reactions resulting in the breakdown of mRNA transcribed in the nucleus.

Table 4.8: Modules identified in each network with six or more nodes and their degree (size). The entries are sorted by the module size.

HNCA:HC		**BL:HC**		**LumA:HC**	
Module	**Size**	**Module**	**Size**	**Module**	**Size**
22	488	80	148	25	222
45	131	42	122	491	135
4	108	73	108	1580	53
69	98	60	70	171	20
114	43	861	15	943	15
48	22	890	10	890	10
5	9	109	8	527	8
30	7	1397	6	323	8
124	7	1533	6	1245	7
11	6	45	6	728	7
		3	6	894	6
				825	6
				853	6
				935	6
				696	6
				163	6
LumB:HC		**HER2:HC**		**NL:HC**	
Module	**Size**	**Module**	**Size**	**Module**	**Size**
71	149	20	214	218	123
48	91	233	120	239	108
17	83	213	72	132	105
63	53	40	51	57	69
1534	15	179	15	94	64
861	15	91	9	88	22
818	15			270	21
128	12			96	12
880	8			58	9
882	8			163	9
890	7			723	7
474	7			135	6
72	7				
293	7				

Chapter 5

Discussion

The two aims of this thesis can be described collectively as the goal of identifying significant pattern changes of differentially expressed genes by using the CSD framework on different breast cancer-related conditions and comparing each of them to healthy individuals. The outcome of the differential gene co-expression analysis applied to different breast cancer-related transcriptomic data presented in Section 4.2 and 4.3 gave insight into important genes and biologically relevant cluster of genes. These included hubs linked to cancer characteristics and modules enriched in processes involved in cancerous behavior.

A central part of network theory is the hierarchical structure, in which hubs participate in many interactions and modules carry out discrete functions [40]. With this in mind it is important to look at the central players and modular structures within the networks as they are likely to have relevant functions. The CSD approach identify three types of differential co-expression patterns that can easily be interpreted in the generated network. It distinguishes between expression pattern of gene pairs that is strongly correlated and maintained between the conditions, expression patterns that are condition-specific, and expression patterns that are opposite between the conditions. Compared to other differential co-expression analysis methods, the CSD provide detection of two types of differential co-expression (S, D) and incorporates the conserved co-expression in the generated network [11].

5.1 Using CSD_R

There is a difference in the calculation of correlation and variance in the original CSD code (CSD_O) and the alternative CSD code (CSD_R) used in this thesis. The CSD_O calculates a pair-wise gene co-expression correlation for all gene expression data points once, followed by variance estimation of sub-samples, in each condition separately, before calculating the C, S and D link score. On the other hand, the CSD_R employ bootstrap resampling, drawing a collection of samples from one condition with each bootstrap iteration and calculating both the correlation and the variance of the gene pairs, before repeating with the other. Ultimately taking the mean of the pair-wise calculations within one conditions, before calculating the C, S, and D scores. The resulting networks are thus based on different approaches but both try to capture meaningful biological complex systems and identify differential gene co-expression changes. Between the two, the CSD_R was opted for as the computational demand of the CSD_O made it too time-consuming.

Before employing the CSD_R on all networks there was a need to determine a basis of the required bootstrap iterations for stable and repeatable results. The requirement is for the estimated values to be stable when comparing them between different iteration numbers, to ensure repeatable results. Running the CSD_R shoved that not only unstable correlation and variance could affect the results, but also a small score interval. The small score interval led to small changes altering the order of gene pairs, providing new results. Comparing 20 and 40 bootstrap iterations of the BL:HC netwoth only produced 41 identical S-linked gene pairs, see Table 4.1. This was also observed with the D-linked gene pairs but not to the same extent. This resulted in a need for higher iteration numbers, gradually improving the concordance between the S-linked gene pairs and highlighted the importance of high iteration number when a link type is observed with a small score interval.

The selection of score was arbitrarily set to the top 1000 of each link type, with the aim of creating a network of a size so that it is meaningful for analysis and of suitable size. As this selection follows no justified selection process, an assessment of the choice was made by inspecting the link distribution of CSD_O networks generated with only 50 of the samples for each of the expression data sets. Resulting in a network constructed with an incomplete data set, only intended for assessing the choice of 1000 of each link type. The degree distribution of the networks, given in Table 4.2, report a similar distribution in all of the generated CSD_O networks. Initially, the expectation was that the links in the HNCA:HC network would be mainly conserved as the comparison is made between similar tissues, but this indicate validity in using the same selection of the 1000 top scores of each link type in the HNCA:HC network generated with HNCS:HC.

5.2 CSD analysis

The networks generated by using the alternative CSD approach were scale-free and far from random, representing complex biological systems. The result show that the networks encompass important cancer mechanisms that is supported by literature. Biological processes enriched in the networks of breast cancer subtypes showed dysregulation of processes involved in cancerous behavior, such as apoptosis, proliferation, motility and immune responses. The network comparing histologically normal tissue and healthy controls on the other hand is not cancerous but encompass previously identified processes of immune and inflammatory response. It is however, important to note that the C, S, and D links do not represent gene regulatory networks or protein-protein interaction (PPI) networks. Consequently, these networks alone can not be used to conclude on any regulatory mechanisms underlying the given phenotype, but rather point to genes that may be involved in these mechanisms.

5.2.1 Application to Histologically Normal Cancer-Adjacent Tissue

The first aim was to identify transcriptional alterations of expression data from histologically normal cancer-adjacent (HNCA) tissue samples, taken outside the tumor margin of breast cancer patients, by comparing it to expression data from healthy controls (HC) of breast tissue without a tumor present. HNCA tissue is often used as control in breast cancer studies, and there have been observations of change in these samples compared to HC samples [23, 25, 24]. This motivates the investigation of comparing differential gene co-expression between HNCA samples and HC to identify genetic interplay that is changed and look at how HNCA breast tissue may be affected by the presence of the tumor. Here, the CSD approach was used as the primary mean to identify significant associations between differentially expressed genes (DEGs) of 144 HNCA samples and 396 HCs samples. Further investigations was looking at central players in the network and identifying modules that may represent functional, and ultimately disease modules.

The use of differential gene co-expression analysis has not, to my knowledge, been employed on HNCA tissue and HC expression data to identify changes in co-expression patterns. The resulting network showed enrichment in processes of lipid metabolism as is expected for breast tissue given its role in lipid and energy homeostasis. Other enriched biological processes are involved in inflammatory response and immune system, indicating a change in cellular behavior between HNCA tissue and HC. As discussed above the inclusion of the same amount of conserved, specific, and differentiated links is not unreasonable, as the original CSD framework result in a similar link distribution. The scale-free topology further validates the non-random, and likely biological relevance of the network.

Many hubs in the network are mainly conserved, indicating multiple maintained interactions for important participants between the assumed normal samples of HNCA breast samples and HC from individuals without a breast tumor present. These processes showed a rich involvement in different aspects of lipid metabolism and energy homeostasis. However, not all hubs are homogeneously conserved, pointing to a changed relation of these hubs dominated by specific or differentiated links. The hubs with links related to an altered co-expression pattern, either specific or differentiated, show changes in transcriptional regulation and gene expression, as well as inflammation. Affirming the initial observation of changed interplay within the network and indication that the tumor may influence the HNCA tissue in some way.

To investigate the presence of functional modules that may represent functionalities that have been altered, modules where identified with Louvain algorithm, producing several modular structures with enriched processes. These further show a changed gene interactions, pointing to possible changes in functional mechanisms. Two of them are dominated by specific links, which indicate a loss or gain of differential co-expression relation, and are involved in regulatory processes of gene expression, and translation of secretory proteins, resulting in changed cellular behavior that is hard to elucidate. The module dominated by differentiated links shows involvement in immune responses.

Collectively these observations point to some changes in co-expression patterns between HNCA breast tissue and HC, involved in gene expression and translation, but also inflammatory and immune responses. Indicating that the HNCA tissue, that is assumed biologically normal due to normal histology, has a changed behavior.

5.2.2 Application to Breast Cancer Intrinsic Subtypes

The second aim was to identify differential gene co-expression patterns that change in the different breast cancer intrinsic subtypes, by comparing expression data from each of the subtypes to the HC expression data. In order to do so the 396 samples of HC were compared to the 718, 489, 240, 330 and 198 samples of the respective subtypes: Luminal A, Luminal B, HER2-enriched, Basal-like and Normal-like. These five intrinsic subtypes are not the only proposed subtypes of breast cancer [27], but the PAM50, clinically used as Prosigna, has been shown to have prognostic and chemotherapeutic response prediction value, separating the breast cancer tumor into clinically relevant subtypes.

Breast cancer is a heterogeneous disease with a variety of malfunctions that can lead to the diseased, cancerous phenotype. Several pathways are associated with cancers obtaining the characteristics needed for their invasive, unimpeded proliferation [134]. A malfunction in any of the genes partaking in a given pathway may result in the same or similar functional change. With the heterogeneity and multitude of genetic alterations that can affect the disease development I do not expected to easily identify clear disease-related co-expression patterns in the CSD networks. However, there is a clear enrichment for processes related to the cancerous phenotype, such as proliferation, apoptosis, immune response, and motility in the networks. This suggest that the networks highlight meaningful relationships between the genes of the breast cancer subtypes and the HC.

Investigating the hubs of the networks and identifying modules revealed relevant functionality and possible contributors to the cancerous behavior. Each of the hubs inhabit functionality that is characteristic of cancer and has a specific or differentiated link dominance. The involvement of these hubs in processes like cell proliferation, fate determination, and energy metabolism emphasize the relevance of the observed changes between a breast cancer intrinsic subtype and HC in the networks. Each of the hubs representing central players that is likely to contribute to the disease phenotype. To associate genes with cancerous characteristics the Louvain algorithm was used to partition the network into communities, and biological process enrichment analysis was performed. Many of the modules were separated so that they contained a common link type. The following investigations was focused on the modules with mainly specific or differentiated links, because this implies gene pairs that correlated oppositely or only in one condition. Modules connected by mainly differentiated or specific links that are enriched in processes like telomere maintenance, growth and immune response are likely candidates to contribute to the disease phenotypes.

There is a clear involvement in processes that facilitate growth and other pro-carcinogenic properties in all of the five networks. Specifically the hubs *TVP23C*, *COA8*, and *CFAP74* in the HER2-enriched subtype and the putative cell cycle gene *CDKN2AIPNL* and *GJC1* in the Normal-like subtype represent possible candidates for further analysis of their respective subtypes, as they, to my knowledge, have not been directly linked to breast cancer.

The findings are of biological relevance and importance, but the heterogeneous nature of the disease can make distinct mechanisms difficult to identify. The intrinsic subtyping, dividing breast cancer based on molecular differences of distinct phenotypes clearly show that phenotypes have molecular characteristics. However, it does not identify the mechanisms of how these phenotypes developed. The underlying mechanisms are many and, as mentioned, can contribute to the same phenotypic alterations, making their pattern difficult to identify with co-expression-related analyses, such as the CSD method. With this in mind, it could also imply that the CSD method, when applied to heterogeneous diseases, capture common features of each of the intrinsic phenotypes. Given the clear association to cancer-related processes, it is likely that the networks are related to mechanisms underlying some aspect of the breast cancer subtypes.

Chapter 6

Conclusion and Outlook

There were two aims of this thesis. The first aim was to use differential gene co-expression analysis to investigate the influence breast tumor had on the surrounding tissue outside the tumor margins, as these are assumed biologically normal and used as control in breast cancer research. The second aim was to use the same analysis to discover genes and modules that are relevant for the disease phenotype of five breast cancer subtypes.

The primary tool for analysis was the CSD framework, supplemented by analyses with enrichment and network tools. The applied method involved alternative calculations of correlation and variance using bootstrapping and existing packages in R, resulting in faster computations. Using this alternative CSD approach (CSD_R), assessment of stability is required. Based on the results a high number of bootstrap iterations is needed when encountering small score intervals. The application of CSD_R with the arbitrary choice of keeping the top 1000 scores of each link type generated networks with similar link distribution to corresponding networks generated with the original CSD calculations (CSD_O). Indicating validity in generating the networks by selecting the top 1000 scores of each link type.

Application of the CSD_R method to compare histologically normal cancer-adjacent (HNCA) tissue samples of the breast with breast tissue samples from healthy individuals, generated a biological relevant network showing a change in cellular behavior between HNCA tissue and healthy controls (HCs). The hubs identified in the network showed conservation in processes of lipid and energy homeostasis but pointed to a change in transcription and gene expression. The modules further supported this change in transcriptional activity and also pointed to changes in processes of the immune system.

More research is needed to understand how HNCA tissue is influenced by the tumor and to what extent the use of HNCA tissue as control in breast cancer studies affect the results. However, it is clear that there is some differences between HNCA tissue and healthy tissue from individual without a breast cancer. It could be interesting to perform CSD analyses using HNCA tissue as the control, accompanied with the same analyses with HC samples as the control, to investigate how HNCA tissue as control may affect the resulting networks. The observed difference in co-expression pattern indicate that genes relevant to breast cancer research may be masked, or discordant, when using HNCA tissue samples as the control.

The CSD_R was also applied to each of the breast cancer intrinsic subtypes with HC and showed to represent biological relevant networks with processes involved in cancerous behavior, such as proliferation and apoptosis. The hub genes were mainly related to a differential co-expression, either specific or differentiated, and inhabited functions important for cancer development and progression, like proliferation, motility and fate determination. Investigation of network modules further support the observed change in cancer related functions, like telomere maintenance and immune response, in the networks.

All networks show a clear involvement in pro-carcinogenic properties and specifically the hub genes *TVP23C*, *COA8*, and *CFAP74* in the HER2-enriched subtype and *CDKN2AIPNL* and *GJC1* in the Normal-like subtype are of interest in for further analysis for their involvement in their respective subtypes. These hubs are linked to cancerous properties, but has to my knowledge, not been directly linked to breast cancer. Additionally, each of the networks contain at least one module that were mostly linked by the differentiated link-type, enriched in processes relevant to the cancerous phenotype. It would have been of further interest to compare the interactions of the networks and of the modules with a protein protein interaction network or metabolic network in order to see how the gene products in the network interact or to explore if there are any metabolic changes improving the growth capacity and other metabolic limited processes like energy production.

Collectively, the use of CSD_R for differential gene co-expression analysis captured interactions of biological relevance and elucidated potential genes and modules involved in the underlying mechanisms of the breast cancer subtypes that could of interest for further studies. The method also highlighted changes in HNCA tissue compared to HC that potentially affect the results of breast cancer research when using HNCA as the control.

Bibliography

[1] Barabási AL. Network Science. Cambridge University Press; 2016. Available from: http://networksciencebook.com/.

[2] Voit EO. A First Course in Systems Biology. 1st ed. New York, NY: Garland Science, Taylor & Francis Group; 2013. ISBN: 978-0-8153-4467-4.

[3] Ratnakumar A, Weinhold N, Mar JC, Riaz N. Protein-Protein interactions uncover candidate 'core genes' within omnigenic disease networks. PLOS Genetics. 2020 jul;16(7):e1008903. Available from: https://doi.org/10.1371/journal.pgen.1008903.

[4] Zhang T, Wang X, Yue Z. Identification of candidate genes related to pancreatic cancer based on analysis of gene co-expression and protein-protein interaction network. Oncotarget; Vol 8, No 41. 2017;Available from: https://doi.org/10.18632/oncotarget.20537.

[5] Salleh SM, Mazzoni G, Løvendahl P, Kadarmideen HN. Gene co-expression networks from RNA sequencing of dairy cattle identifies genes and pathways affecting feed efficiency. BMC Bioinformatics. 2018;19(1):513. Available from: https://doi.org/10.1186/s12859-018-2553-z.

[6] Alon U. An Introduction to systems biology : design principles of biological circuits. vol. 10 of Chapman and Hall/CRC mathematical & computational biology series. Boca Raton, Fla: Chapman & Hall; 2007. ISBN: 9781584886426.

[7] Nopoulos PC. Huntington disease: a single-gene degenerative disorder of the striatum. Dialogues in clinical neuroscience. 2016 mar;18(1):91–98. Available from: https://doi.org/10.31887/DCNS.2016.18.1/pnopoulos.

[8] Aoki K, Ogata Y, Shibata D. Approaches for Extracting Practical Information from Gene Co-expression Networks in Plant Biology. Plant and Cell Physiology. 2007 mar;48(3):381–390. Available from: https://doi.org/10.1093/pcp/pcm013.

[9] Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, et al. Variations in DNA elucidate molecular networks that cause disease. Nature. 2008;452(7186):429–435. Available from: https://doi.org/10.1038/nature06757.

[10] Yang Y, Han L, Yuan Y, Li J, Hei N, Liang H. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. Nature Communications. 2014;5(1):3231. Available from: https://doi.org/10.1038/ncomms4231.

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

[12] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research. 2002 jan;30(1):207–210. Available from: https://doi.org/10.1093/nar/30.1.207.

[13] Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, et al. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. Biopreservation and Biobanking. 2015 oct;13(5):311–319. Available from: https://doi.org/10.1089/bio.2015.0032.

[14] Hanahan D, Weinberg RA. The Hallmarks of Cancer. Cell. 2000;100(1):57–70. Available from: https://doi.org/10.1016/S0092-8674(00)81683-9.

[15] Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell. 2011;144(5):646–674. Available from: https://doi.org/10.1016/j.cell.2011.02.013.

[16] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians. 2018;68(6):394–424. Available from: https://doi.org/10.3322/caac.21492.

[17] Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747–752. Available from: https://doi.org/10.1038/35021093.

[18] Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences of the United States of America. 2001 sep;98(19):10869–10874. Available from: https://doi.org/10.1073/pnas.191367098.

[19] Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. Journal of Clinical Oncology. 2009 mar;27(8):1160–1167. Available from: https://doi.org/10.1200/JCO.2008.18.1370.

[20] Giuliano AE, Connolly JL, Edge SB, Mittendorf EA, Rugo HS, Solin LJ, et al. Breast Cancer-Major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. CA: a cancer journal for clinicians. 2017 jul;67(4):290–303. Available from: https://doi.org/10.3322/caac.21393.

[21] Harris LN, Ismaila N, McShane LM, Andre F, Collyar DE, Gonzalez-Angulo AM, et al. Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2016 apr;34(10):1134–1150. Available from: https://doi.org/10.1200/JCO.2015.65.2289.

[22] Duffy MJ, Harbeck N, Nap M, Molina R, Nicolini A, Senkus E, et al. Clinical use of biomarkers in breast cancer: Updated guidelines from the European Group on Tumor Markers (EGTM). European Journal of Cancer. 2017;75:284–298. Available from: https://doi.org/10.1016/j.ejca.2017.01.017.

[23] Aran D, Camarda R, Odegaard J, Paik H, Oskotsky B, Krings G, et al. Comprehensive analysis of normal adjacent to tumor transcriptomes. Nature Communications. 2017;8(1):1077. Available from: https://doi.org/10.1038/s41467-017-01027-z.

[24] Troester MA, Hoadley KA, D'Arcy M, Cherniack AD, Stewart C, Koboldt DC, et al. DNA defects, epigenetics, and gene expression in cancer-adjacent breast: a study from The Cancer Genome Atlas. NPJ breast cancer. 2016;2:16007. Available from: https://doi.org/10.1038/npjbcancer.2016.7.

[25] Casbas-Hernandez P, Sun X, Roman-Perez E, D'Arcy M, Sandhu R, Hishida A, et al. Tumor intrinsic subtype is reflected in cancer-adjacent tissue. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2015 feb;24(2):406–414. Available from: https://doi.org/10.1158/1055-9965.EPI-14-0934.

[26] Weigelt B, Baehner FL, Reis-Filho JS. The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. The Journal of Pathology. 2010 jan;220(2):263–280. Available from: https://doi.org/10.1002/path.2648.

[27] Alizart M, Saunus J, Cummings M, Lakhani SR. Molecular classification of breast carcinoma. Diagnostic Histopathology. 2012;18(3):97–103. Available from: https://doi.org/10.1016/j.mpdhp.2011.12.003.

[28] Harbeck N, Penault-Llorca F, Cortes J, Gnant M, Houssami N, Poortmans P, et al. Breast cancer. Nature Reviews Disease Primers. 2019;5(1):66. Available from: https://doi.org/10.1038/s41572-019-0111-2.

[29] Feng Y, Spezia M, Huang S, Yuan C, Zeng Z, Zhang L, et al. Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. Genes & diseases. 2018 may;5(2):77–106. Available from: https://doi.org/10.1016/j.gendis.2018.05.001.

[30] Sørlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proceedings of the National Academy of Sciences. 2003 jul;100(14):8418 LP – 8423. Available from: https://doi.org/10.1073/pnas.0932692100.

[31] Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences of the United States of America. 2002 may;99(10):6567–6572. Available from: https://doi.org/10.1073/pnas.082099299.

[32] Guedj M, Marisa L, de Reynies A, Orsetti B, Schiappa R, Bibeau F, et al. A refined molecular taxonomy of breast cancer. Oncogene. 2012;31(9):1196–1206. Available from: https://doi.org/10.1038/onc.2011.301.

[33] Bastien RRL, Rodríguez-Lescure Á, Ebbert MTW, Prat A, Munárriz B, Rowe L, et al. PAM50 breast cancer subtyping by RT-qPCR and concordance with standard clinical molecular markers. BMC medical genomics. 2012 oct;5:44. Available from: https://doi.org/10.1186/1755-8794-5-44.

[34] Gnant M, Filipits M, Greil R, Stoeger H, Rudas M, Bago-Horvath Z, et al. Predicting distant recurrence in receptor-positive breast cancer patients with limited clinicopathological risk: using the PAM50 Risk of Recurrence score in 1478 postmenopausal patients of the ABCSG-8 trial treated with adjuvant endocrine therapy alone. Annals of Oncology. 2014;25(2):339–345. Available from: https://doi.org/10.1093/annonc/mdt494.

[35] SLAUGHTER DP, SOUTHWICK HW, SMEJKAL W. Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin. Cancer. 1953 sep;6(5):963–968. Available from: https://doi.org/10.1002/1097-0142(195309)6:5<963::aid-cncr2820060515>3.0.co;2-q.

[36] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews Genetics. 2009 jan;10(1):57–63. Available from: https://doi.org/10.1038/nrg2484.

[37] Cammack R, Atwood T, Campbell P, Parish H, Smith A, Vella F, et al.. gene expression profile. Oxford University Press; 2008. Available from: https://doi.org/10.1093/acref/9780198529170.013.7770.

[38] McDermaid A, Monier B, Zhao J, Liu B, Ma Q. Interpretation of differential gene expression results of RNA-seq data: review and integration. Briefings in Bioinformatics. 2018;20(6):2044–2054. Available from: https://doi.org/10.1093/bib/bby067.

[39] Mendoza MLZ, Resendis-Antonio O. Modules, Identification Methods and Biological Function. In: Dubitzky W, Wolkenhauer O, Cho KH, Yokota H, editors. Encyclopedia of Systems Biology. New York, NY: Springer New York; 2013. p. 1450–1453. Available from: https://doi.org/10.1007/978-1-4419-9863-7_1315.

[40] Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL. Hierarchical Organization of Modularity in Metabolic Networks. Science. 2002 aug;297(5586):1551 LP – 1555. Available from: https://doi.org/10.1126/science.1073374.

[41] Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nature Reviews Genetics. 2011;12(1):56–68. Available from: https://doi.org/10.1038/nrg2918.

[42] Voigt A, Almaas E. Assessment of weighted topological overlap (wTO) to improve fidelity of gene co-expression networks. BMC Bioinformatics. 2019;20(1):58. Available from: https://doi.org/10.1186/s12859-019-2596-9.

[43] Nowick K, Gernat T, Almaas E, Stubbs L. Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. Proceedings of the National Academy of Sciences. 2009 dec;106(52):22358 LP – 22363. Available from: https://doi.org/10.1073/pnas.0911376106.

[44] Correlation Coefficient. In: The Concise Encyclopedia of Statistics. New York, NY: Springer New York; 2008. p. 115–119. Available from: https://doi.org/10.1007/978-0-387-32833-1_83.

[45] Dormann C. Correlation and Association. In: Enviromental Data Analysis. Springer, Cham; 2020. p. 65–70. Available from: https://doi.org/10.1007/978-3-030-55020-2_5.

[46] Henderson AR. The bootstrap: A technique for data-driven statistics. Using computer-intensive analyses to explore experimental data. Clinica Chimica Acta. 2005;359(1):1–26. Available from: https://doi.org/10.1016/j.cccn.2005.04.002.

[47] Chernick MR. Bootstrap Methods A Guide for Practitioners and Researchers. 2nd ed. Hoboken: John Wiley & Sons; 2011. ISBN: 978-1-118-21159-5.

[48] Booth JG, Sarkar S. Monte Carlo Approximation of Bootstrap Variances. The American Statistician. 1998 dec;52(4):354–357. Available from: https://doi.org/10.2307/2685441.

[49] Skelly AC, Dettori JR, Brodt ED. Assessing bias: the importance of considering confounding. Evidence-based spine-care journal. 2012 feb;3(1):9–12. Available from: https://doi.org/10.1055/s-0031-1298595.

[50] confounding. In: Hine R, editor. A Dictionary of Biology. 8th ed. Oxford University Press; 2019. Available from: https://www.oxfordreference.com/view/10.1093/acref/9780198821489.001.0001/acref-9780198821489-e-4801.

[51] Hypothesis Testing. In: The Concise Encyclopedia of Statistics. New York, NY: Springer New York; 2008. p. 250–252. Available from: https://doi.org/10.1007/978-0-387-32833-1_184.

[52] Banerjee A, Chitnis UB, Jadhav SL, Bhawalkar JS, Chaudhury S. Hypothesis testing, type I and type II errors. Industrial psychiatry journal. 2009 jul;18(2):127–131. Available from: https://doi.org/10.4103/0972-6748.62274.

[53] Higdon R. Multiple Hypothesis Testing. In: Dubitzky W, Wolkenhauer O, Cho KH, Yokota H, editors. Encyclopedia of Systems Biology. New York, NY: Springer New York; 2013. p. 1468–1469. Available from: https://doi.org/10.1007/978-1-4419-9863-7_1211.

[54] Dudoit S, Shaffer JP, Boldrick JC. Multiple Hypothesis Testing in Microarray Experiments. Statistical Science. 2003 dec;18(1):71–103. Available from: http://www.jstor.org/stable/3182872.

[55] Haynes W. Bonferroni Correction. In: Dubitzky W, Wolkenhauer O, Cho KH, Yokota H, editors. Encyclopedia of Systems Biology. New York, NY: Springer New York; 2013. p. 154. Available from: https://doi.org/10.1007/978-1-4419-9863-7_1213.

[56] Rouam S. False Discovery Rate (FDR). In: Dubitzky W, Wolkenhauer O, Cho KH, Yokota H, editors. Encyclopedia of Systems Biology. New York, NY: Springer New York; 2013. p. 731–732. Available from: https://doi.org/10.1007/978-1-4419-9863-7_223.

[57] Serres MH, Gopal S, Nahum LA, Liang P, Gaasterland T, Riley M. A functional update of the Escherichia coli K-12 genome. Genome biology. 2001;2(9):RESEARCH0035–RESEARCH0035. Available from: https://doi.org/10.1186/gb-2001-2-9-research0035.

[58] Jorde LB, Carey JC, Bamshad MJ. Medical Genetics. 5th ed. Elsevier; 2016. ISBN: 978-0-323-18835-7.

[59] van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. Gene co-expression analysis for functional classification and gene-disease predictions. Briefings in Bioinformatics. 2018 jul;19(4):575–592. Available from: https://doi.org/10.1093/bib/bbw139.

[60] de la Fuente A. From 'differential expression' to 'differential networking' – identification of dysfunctional regulatory networks in diseases. Trends in Genetics. 2010;26(7):326–333. Available from: https://doi.org/10.1016/j.tig.2010.05.001.

[61] Choi JK, Yu U, Yoo OJ, Kim S. Differential coexpression analysis using microarray data and its application to human cancer. Bioinformatics. 2005 dec;21(24):4348–4355. Available from: https://doi.org/10.1093/bioinformatics/bti722.

[62] Liu BH, Yu H, Tu K, Li C, Li YX, Li YY. DCGL: an R package for identifying differentially coexpressed genes and links from gene expression microarray data. Bioinformatics. 2010 oct;26(20):2637–2638. Available from: https://doi.org/10.1093/bioinformatics/btq471.

[63] Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nature Genetics. 2013;45(6):580–585. Available from: https://doi.org/10.1038/ng.2653.

[64] Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012 apr;486:346. Available from: https://doi.org/10.1038/nature10983.

[65] Morselli Gysi D, de Miranda Fragoso T, Zebardast F, Bertoli W, Busskamp V, Almaas E, et al. Whole transcriptomic network analysis using Co-expression Differential Network Analysis (CoDiNA). PLOS ONE. 2020 oct;15(10):e0240523. Available from: https://doi.org/10.1371/journal.pone.0240523.

[66] Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome research. 2003 sep;13(9):2129–2141. Available from: https://doi.org/10.1101/gr.772403.

[67] Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Research. 2007 jul;35(suppl_2):W169–W175. Available from: https://doi.org/10.1093/nar/gkm415.

[68] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nature Genetics. 2000;25(1):25–29. Available from: https://doi.org/10.1038/75556.

[69] Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research. 2009 jan;37(1):1–13. Available from: https://doi.org/10.1093/nar/gkn923.

[70] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 2008;2008(10):P10008. Available from: http://dx.doi.org/10.1088/1742-5468/2008/10/P10008.

[71] Darshi M, Mendiola VL, Mackey MR, Murphy AN, Koller A, Perkins GA, et al. ChChd3, an inner mitochondrial membrane protein, is essential for maintaining crista integrity and mitochondrial function. The Journal of biological chemistry. 2011 jan;286(4):2918–2932. Available from: https://doi.org/10.1074/jbc.M110.171975.

[72] Collins T, Sander TL. The Superfamily of SCAN Domain Containing Zinc Finger Transcription Factors. In: Madame Curie Bioscience Database [Internet]. Austin (TX): Landes Bioscience; 2000-2013. Available from: https://www.ncbi.nlm.nih.gov/books/NBK6264/.

[73] Saci A, Cantley LC, Carpenter CL. Rac1 regulates the activity of mTORC1 and mTORC2 and controls cellular size. Molecular cell. 2011 apr;42(1):50–61. Available from: https://doi.org/10.1016/j.molcel.2011.03.017.

[74] Marei H, Malliri A. Rac1 in human diseases: The therapeutic potential of targeting Rac1 signaling regulatory mechanisms. Small GTPases. 2017 jul;8(3):139–163. Available from: https://doi.org/10.1080/21541248.2016.1211398.

[75] Qin C, Liu R, Liu H. The conflicting role of Rac1 in inflammation. Inflammation and Cell Signaling. 2015;2(3).

[76] Davey RJ, Moens PD. Profilin: many facets of a small protein. Biophysical reviews. 2020 aug;12(4):827–849. Available from: https://doi.org/10.1007/s12551-020-00723-3.

[77] Ni Y, Ji C, Wang B, Qiu J, Wang J, Guo X. A Novel pro-adipogenesis factor abundant in adipose tissues and over-expressed in obesity acts upstream of PPARγ and C/EBPα. Journal of Bioenergetics and Biomembranes. 2013;45(3):219–228. Available from: https://doi.org/10.1007/s10863-012-9492-6.

[78] Gao Y, Zhang X, Wang T, Zhang Y, Wang Q, Hu Y. HNRNPCL1, PRAMEF1, CFAP74, and DFFB: Common Potential Biomarkers for Sporadic and Suspected Lynch Syndrome Endometrial Cancer. Cancer management and research. 2020 nov;12:11231–11241. Available from: https://doi.org/10.2147/CMAR.S262421.

[79] Sasajima Y, Tanaka H, Miyake S, Yuasa Y. A novel EID family member, EID-3, inhibits differentiation and forms a homodimer or heterodimer with EID-2. Biochemical and Biophysical Research Communications. 2005;333(3):969–975. Available from: https://doi.org/10.1016/j.bbrc.2005.06.013.

[80] Stenson Holst L, Langin D, Mulder H, Laurell H, Grober J, Bergh A, et al. Molecular Cloning, Genomic Organization, and Expression of a Testicular Isoform of Hormone-Sensitive Lipase. Genomics. 1996;35(3):441–447. Available from: https://doi.org/10.1006/geno.1996.0383.

[81] Basel-Vanagaite L, Zevit N, Zahav AH, Guo L, Parathath S, Pasmanik-Chor M, et al. Transient Infantile Hypertriglyceridemia, Fatty Liver, and Hepatic Fibrosis Caused by Mutated GPD1, Encoding Glycerol-3-Phosphate Dehydrogenase 1. The American Journal of Human Genetics. 2012;90(1):49–60. Available from: https://doi.org/10.1016/j.ajhg.2011.11.028.

[82] Tansey JT, Sztalryd C, Hlavin EM, Kimmel AR, Londos C. The central role of perilipin a in lipid metabolism and adipocyte lipolysis. IUBMB life. 2004 jul;56(7):379–385. Available from: https://doi.org/10.1080/15216540400009968.

[83] Ceperuelo-Mallafrcé V, Miranda M, Chacón MR, Vilarrasa N, Megia A, Gutiérrez C, et al. Adipose Tissue Expression of the Glycerol Channel Aquaporin-7 Gene Is Altered in Severe Obesity But Not in Type 2 Diabetes. The Journal of Clinical Endocrinology & Metabolism. 2007 sep;92(9):3640–3645. Available from: https://doi.org/10.1210/jc.2007-0531.

[84] Duan X, Krycer JR, Cooke KC, Yang G, James DE, Fazakerley DJ. Membrane Topology of Trafficking Regulator of GLUT4 1 (TRARG1). Biochemistry. 2018 jul;57(26):3606–3615. Available from: https://doi.org/10.1021/acs.biochem.8b00361.

[85] Puri V, Konda S, Ranjit S, Aouadi M, Chawla A, Chouinard M, et al. Fat-specific Protein 27, a Novel Lipid Droplet Protein That Enhances Triglyceride Storage. Journal of Biological Chemistry. 2007;282(47):34213–34218. Available from: https://doi.org/10.1074/jbc.M707404200.

[86] Llamas A, Chamizo-Ampudia A, Tejada-Jimenez M, Galvan A, Fernandez E. The molybdenum cofactor enzyme mARC: Moonlighting or promiscuous enzyme? BioFactors (Oxford, England). 2017 jul;43(4):486–494. Available from: https://doi.org/10.1002/biof.1362.

[87] Robbins AL, Savage DB. The genetics of lipid storage and human lipodystrophies. Trends in Molecular Medicine. 2015;21(7):433–438. Available from: https://doi.org/10.1016/j.molmed.2015.04.004.

[88] Tsakogiannis D, Kalogera E, Zagouri F, Zografos E, Balalis D, Bletsa G. Determination of FABP4, RBP4 and the MMP-9/NGAL complex in the serum of women with breast cancer. Oncol Lett. 2021;21(2):85. Available from: https://doi.org/10.3892/ol.2020.12346.

[89] Micello D, Bossi A, Marando A, Dainese E, Sessa F, Capella C. Expression of calretinin in high-grade hormone receptor-negative invasive breast carcinomas: correlation with histological and molecular subtypes. Virchows Archiv. 2017;471(1):13–21. Available from: https://doi.org/10.1007/s00428-017-2149-4.

[90] Powell G, Roche H, Roche WR. Expression of calretinin by breast carcinoma and the potential for misdiagnosis of mesothelioma. Histopathology. 2011 nov;59(5):950–956. Available from: https://doi.org/10.1111/j.1365-2559.2011.04031.x.

[91] Mu J, Roach PJ. Characterization of Human Glycogenin-2, a Self-glucosylating Initiator of Liver Glycogen Metabolism. Journal of Biological Chemistry. 1998;273(52):34850–34856. Available from: https://doi.org/10.1074/jbc.273.52.34850.

[92] Kaptain S, Downey WE, Tang C, Philpott C, Haile D, Orloff DG, et al. A regulated RNA binding protein also possesses aconitase activity. Proceedings of the National Academy of Sciences of the United States of America. 1991 nov;88(22):10109–10113. Available from: https://doi.org/10.1073/pnas.88.22.10109.

[93] Eisenstein RS. IRON REGULATORY PROTEINS AND THE MOLECULAR CONTROL OF MAMMALIAN IRON METABOLISM. Annual Review of Nutrition. 2000 jul;20(1):627–662. Available from: https://doi.org/10.1146/annurev.nutr.20.1.627.

[94] Pabst T, Mueller BU, Zhang P, Radomska HS, Narravula S, Schnittger S, et al. Dominant-negative mutations of CEBPA, encoding CCAAT/enhancer binding protein-α (C/EBPα), in acute myeloid leukemia. Nature Genetics. 2001;27(3):263–270. Available from: https://doi.org/10.1038/85820.

[95] Kothari C, Diorio C, Durocher F. The Importance of Breast Adipose Tissue in Breast Cancer. International journal of molecular sciences. 2020 aug;21(16):5760. Available from: https://doi.org/10.3390/ijms21165760.

[96] Parimisetty A, Dorsemans AC, Awada R, Ravanan P, Diotel N, Lefebvre d'Hellencourt C. Secret talk between adipose tissue and central nervous system via secreted factors—an emerging frontier in the neurodegenerative research. Journal of Neuroinflammation. 2016;13(1):67. Available from: https://doi.org/10.1186/s12974-016-0530-x.

[97] Garcia-Barcena C, Osinalde N, Ramirez J, Mayor U. How to Inactivate Human Ubiquitin E3 Ligases by Mutation ; 2020. Available from: https://www.frontiersin.org/article/10.3389/fcell.2020.00039.

[98] Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell JE. Protein Sorting: Organelle Biogenesis and Protein Secretion. In: Molecular Cell Biology. 4th ed. New York: W.H. Freeman;. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21475/.

[99] Dendrite. In: Binder MD, Hirokawa N, Windhorst U, editors. Encyclopedia of Neuroscience. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009. p. 936. Available from: https://doi.org/10.1007/978-3-540-29678-2_1439.

[100] Dong Y, Zhang L, Zhang S, Bai Y, Chen H, Sun X, et al. Phosphatase of Regenerating Liver 2 (PRL2) Is Essential for Placental Development by Down-regulating PTEN (Phosphatase and Tensin Homologue Deleted on Chromosome 10) and Activating Akt Protein. Journal of Biological Chemistry. 2012;287(38):32172–32179. Available from: https://doi.org/10.1074/jbc.M112.393462.

[101] Hemmings BA, Restuccia DF. PI3K-PKB/Akt pathway. Cold Spring Harbor perspectives in biology. 2012 sep;4(9):a011189–a011189. Available from: https://doi.org/10.1101/cshperspect.a011189.

[102] Fuller-Pace FV. DEAD box RNA helicase functions in cancer. RNA biology. 2013 jan;10(1):121–132. Available from: https://doi.org/10.4161/rna.23312.

[103] Wortham NC, Ahamed E, Nicol SM, Thomas RS, Periyasamy M, Jiang J, et al. The DEAD-box protein p72 regulates ERalpha-/oestrogen-dependent transcription and cell growth, and is associated with improved survival in ERalpha-positive breast cancer. Oncogene. 2009 nov;28(46):4053–4064. Available from: https://doi.org/10.1038/onc.2009.261.

[104] Holm M, Joenväärä S, Saraswat M, Mustonen H, Tohmola T, Ristimäki A, et al. Identification of several plasma proteins whose levels in colorectal cancer patients differ depending on outcome. FASEB BioAdvances. 2019 dec;1(12):723–730. Available from: https://doi.org/10.1096/fba.2019-00062.

[105] Inadome H, Noda Y, Kamimura Y, Adachi H, Yoda K. Tvp38, Tvp23, Tvp18 and Tvp15: Novel membrane proteins in the Tlg2-containing Golgi/endosome compartments of *Saccharomyces cerevisiae*. Experimental Cell Research. 2007;313(4):688–697. Available from: https://doi.org/10.1016/j.yexcr.2006.11.008.

[106] Küry S, Besnard T, Ebstein F, Khan TN, Gambin T, Douglas J, et al. De Novo Disruption of the Proteasome Regulatory Subunit *PSMD12* Causes a Syndromic Neurodevelopmental Disorder. The American Journal of Human Genetics. 2017;100(2):352–363. Available from: https://doi.org/10.1016/j.ajhg.2017.01.003.

[107] Signes A, Cerutti R, Dickson AS, Benincá C, Hinchy EC, Ghezzi D, et al. APOPT1/COA8 assists COX assembly and is oppositely regulated by UPS and ROS. EMBO Molecular Medicine. 2019 jan;11(1):e9582. Available from: https://doi.org/10.15252/emmm.201809582.

[108] Wettschureck N. G-Protein-Coupled Receptors. In: Offermanns S, Rosenthal W, editors. Encyclopedia of Molecular Pharmacology. Cham: Springer International Publishing; 2020. p. 1–9. Available from: https://doi.org/10.1007/978-3-030-21573-6_70-1.

[109] Rourke JL, Muruganandan S, Dranse HJ, McMullen NM, Sinal CJ. Gpr1 is an active chemerin receptor influencing glucose homeostasis in obese mice. The

Journal of endocrinology. 2014 aug;222(2):201–215. Available from: https://doi.org/10.1530/JOE-14-0069.

[110] Huang C, Dai XY, Cai JX, Chen J, Wang BB, Zhu W, et al. A Screened GPR1 Peptide Exerts Antitumor Effects on Triple-Negative Breast Cancer. Molecular Therapy - Oncolytics. 2020;18:602–612. Available from: https://doi.org/10.1016/j.omto.2020.08.013.

[111] Kanter HL, Saffitz JE, Beyer EC. Molecular Cloning of Two Human Cardiac Gap Junction Proteins, Connexin40 and Connexin45. Journal of Molecular and Cellular Cardiology. 1994;26(7):861–868. Available from: https://doi.org/10.1006/jmcc.1994.1103.

[112] Aasen T, Leithe E, Graham SV, Kameritsch P, Mayán MD, Mesnil M, et al. Connexins in cancer: bridging the gap to the clinic. Oncogene. 2019;38(23):4429–4451. Available from: https://doi.org/10.1038/s41388-019-0741-6.

[113] Giotti B, Chen SH, Barnett MW, Regan T, Ly T, Wiemann S, et al. Assembly of a parts list of the human mitotic cell cycle machinery. Journal of Molecular Cell Biology. 2019 aug;11(8):703–718. Available from: https://doi.org/10.1093/jmcb/mjy063.

[114] Matsuo S, Hiasa M, Omote H. Functional characterization and tissue localization of the facilitative glucose transporter GLUT12. The Journal of Biochemistry. 2020 dec;168(6):611–620. Available from: https://doi.org/10.1093/jb/mvaa090.

[115] Rogers S, Docherty SE, Slavin JL, Henderson MA, Best JD. Differential expression of GLUT12 in breast cancer and normal breast tissue. Cancer Letters. 2003;193(2):225–233. Available from: https://doi.org/10.1016/S0304-3835(03)00010-7.

[116] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics. 2000 may;25(1):25–29. Available from: https://doi.org/10.1038/75556.

[117] The Gene Ontology resource: enriching a GOld mine. Nucleic acids research. 2021 jan;49(D1):D325–D334. Available from: https://doi.org/10.1093/nar/gkaa1113.

[118] Kaminski NE, Sulentic CEW. B Lymphocytes. In: Vohr HW, editor. Encyclopedia of Immunotoxicology. Berlin, Heidelberg: Springer Berlin Heidelberg; 2016. p. 105–107. Available from: https://doi.org/10.1007/978-3-642-54596-2_158.

[119] Complement Cascade. In: Vohr HW, editor. Encyclopedia of Immunotoxicology. Berlin, Heidelberg: Springer Berlin Heidelberg; 2016. p. 198. Available from: https://doi.org/10.1007/978-3-642-54596-2_200286.

[120] Freund A, Zhong FL, Venteicher AS, Meng Z, Veenstra TD, Frydman J, et al. Proteostatic control of telomerase function through TRiC-mediated folding of TCAB1. Cell. 2014 dec;159(6):1389–1403. Available from: https://doi.org/10.1016/j.cell.2014.10.059.

[121] Laiosa M, Silverstone A. Thymus: A Mediator of T-Cell Development and Potential Target of Toxicological Agents. In: Vohr HW, editor. Encyclopedia of Immunotoxicology. Berlin, Heidelberg: Springer Berlin Heidelberg; 2016. p. 891–895. Available from: https://doi.org/10.1007/978-3-642-54596-2_1473.

[122] Brasier AR. The NF-κB regulatory network. Cardiovascular Toxicology. 2006;6(2):111–130. Available from: https://doi.org/10.1385/CT:6:2:111.

[123] Sotgia F, Rui H, Bonuccelli G, Mercier I, Pestell RG, Lisanti MP. Caveolin-1, Mammary Stem Cells, and Estrogen-Dependent Breast Cancers. Cancer Research. 2006 nov;66(22):10647 LP – 10651. Available from: https://doi.org/10.1158/0008-5472.CAN-06-2805.

[124] Xue M, Zhang K, Mu K, Xu J, Yang H, Liu Y, et al. Regulation of estrogen signaling and breast cancer proliferation by an ubiquitin ligase TRIM56. Oncogenesis. 2019;8(5):30. Available from: https://doi.org/10.1038/s41389-019-0139-x.

[125] Kolte D, McClung JA, Aronow WS. Vasculogenesis and Angiogenesis. In: Aronow WS, McClung JABT, editors. Translational Research in Coronary Artery Disease. Boston: Academic Press; 2016. p. 49–65. Available from: https://doi.org/10.1016/B978-0-12-802385-3.00006-1.

[126] Ridge SM, Sullivan FJ, Glynn SA. Mesenchymal stem cells: key players in cancer progression. Molecular Cancer. 2017;16(1):31. Available from: https://doi.org/10.1186/s12943-017-0597-8.

[127] Kameritsch P, Renkawitz J. Principles of Leukocyte Migration Strategies. Trends in Cell Biology. 2020 oct;30(10):818–832. Available from: https://doi.org/10.1016/j.tcb.2020.06.007.

[128] Worth RG, Screiber AD. Fc Receptor Phagocytosis. In: Molecular Mechanisms of Phagocytosis. Springer, Boston, MA; 2005. Available from: https://doi.org/10.1007/978-0-387-28669-3_3.

[129] Harjunpää H, Llort Asens M, Guenther C, Fagerholm SC. Cell Adhesion Molecules and Their Roles and Regulation in the Immune and Tumor Microenvironment; 2019. Available from: https://www.frontiersin.org/article/10.3389/fimmu.2019.01078.

[130] Egholm C, Heeb LEM, Impellizzieri D, Boyman O. The Regulatory Effects of Interleukin-4 Receptor Signaling on Neutrophils in Type 2 Immune Responses. Frontiers in Immunology. 2019;10:2507. Available from: https://doi.org/10.3389/fimmu.2019.02507.

[131] Adaptive Immunity. In: Vohr HW, editor. Encyclopedia of Immunotoxicology. Berlin, Heidelberg: Springer Berlin Heidelberg; 2016. p. 8. Available from: https://doi.org/10.1007/978-3-642-54596-2_200018.

[132] Brychta RJ, Chen KY. Cold-induced thermogenesis in humans. European journal of clinical nutrition. 2017 mar;71(3):345–352. Available from: https://doi.org/10.1038/ejcn.2016.223.

[133] Zabel MK, Kirsch WM. From development to dysfunction: microglia and the complement cascade in CNS homeostasis. Ageing research reviews. 2013 jun;12(3):749–756. Available from: https://doi.org/10.1016/j.arr.2013.02.001.

[134] Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nature Medicine. 2004;10(8):789–799. Available from: https://doi.org/10.1038/nm1087.

Appendices

A.1 PAM50 classifier genes

The genes used in the PAM50 classifier to classify the intrinsic subtypes are listed below (Table A.1) accompanied by a heatmap of their relative expression in each of the given subtypes (Figure A.1).

Table A.1: The 50 genes that make up the PAM50 classifier of the intrinsic subtypes developed by Parker et al. [19].

UBE2C	*CEP55*	*KIF2C*	*MMP11*	*NAT1*	*KRT17*	*FGFR4*	*EGFR*
PTTG1	*KNTC2*	*EXO1*	*GRB7*	*GPR160*	*KRT5*	*MYC*	
MYBL2	*UBE2T*	*CDCA1*	*ERBB2*	*FOXA1*	*SFRP1*	*MIA*	
BIRC5	*RRM2*	*CENPF*	*TMEM45B*	*BLVRA*	*BCL2*	*FOXC1*	
CCNB1	*CDC6*	*CCNE1*	*BAG1*	*CXXC5*	*KRT14*	*ACTR3B*	
TYMS	*ANLN*	*MKI67*	*PGR*	*ESR1*	*MLPH*	*PHGDH*	
MELK	*ORC6L*	*CDC20*	*MAPT*	*SLC39A6*	*MDM2*	*CDH3*	

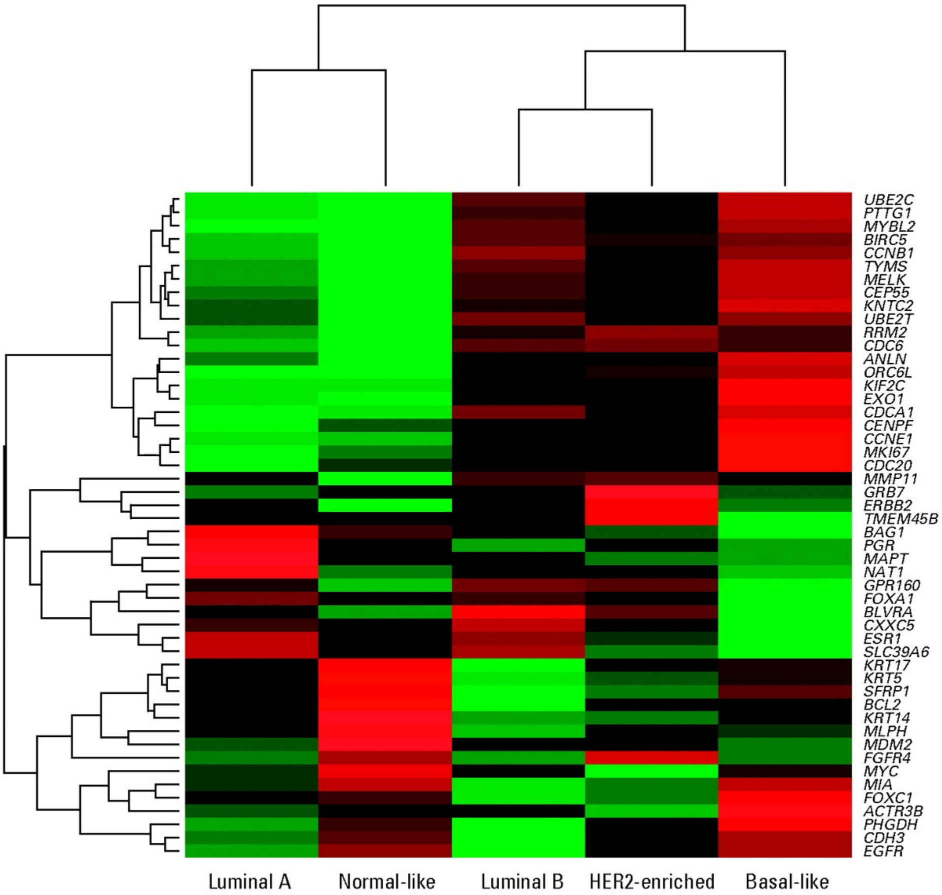


Figure A.1: Heatmap of the PAM50 genes in each of the intrinsic subtypes, shown as red/green according to their relative gene expression level. Source: Figure A3 [19]

A.2 Characteristics of histologically normal cancer-adjacent tissue

Aran et al. revealed altered pathways in across different HNCA tissue types and characterized 18 genes specifically activated in at least three of the cancer-adjacent tissue types, listed in Table A.2 [23]. Pathways related to inflammatory response are generally enriched in the HNCA tissues, and there were also enrichment of some cancer-related processes like apoptosis.

Table A.2: The 18 genes specifically activated in cancer-adjacent tissue of at least three tissue types, available in Supplementary Figure 19 [23].

ACE, *ATF3*, *CNN1*, *CSRP1*, *CXCL12*, *CYR61*, *DPT*, *EGR1*, *EGR2*, *EGR3*, *FGL2*, *FOS*, *FOSB*, *JUND*, *MYADM*, *NR4A3*, *RCAN1* and *TPPP*

A.3 R Code for CSD Calculations and Filtering

The following R code were used to import data sets, run the CSD analysis, and extract the 1000 highest scoring links of each interaction type C, S and D, for each of the condition:control tissue samples (HNCA:HC, BL:HC, LumA:HC, LumB:HC, HER2:HC, and NL:HC). Accordingly, the appropriate changes were made in input filename of the condition, the number of iterations, seed number, and output file names for each run. The script takes two gene expression data sets as input, with the first column containing gene identifiers, the header row containing sample names, and the remaining rows and columns containing the gene expression of each gene.

R Code for running CSD and Filtering Results

```
#!/usr/bin/env Rscript

#############################################################
#
# This script takes two gene expression data sets as input,
# transposing them and sourcing the CSD computation
# and log_progress, available from
# https://github.com/AlmaasLab/csdR to
# calculate C, S and D scores for the gene pairs between the
# conditions.
# This require the file welford.cpp, from the same source
#
# This script require the script find_rho_and_var.R,
# the R packages WGCNA, outparse, glue and magrittr.
#
# Parameters to be filled in by the user:
# Filename of input files in x_1 and x_2
# Seed number (set.seed())
# Number of iterations (n_it)
# Number of links selected of each type (pairs_to_pick)
# Filename of the four .txt outputs of link values
# Filename of the six .jpeg outputs of histograms
#
#############################################################

#Source the CSD function and log_progress, available from
    https://github.com/AlmaasLab/csdR
source("find_rho_and_var.R")

#Import and transpose input files
log_progress("Starting␣to␣read␣files")
x_1 <- read.table("GTEx_norm.txt", header = TRUE, row.names = 1)
    %>% as.matrix() %>% t()
x_2 <- read.table("Her2_metabric.txt", header = TRUE , row.names
    = 1)%>%as.matrix()%>%t()
```

```
log_progress("Files␣imported␣and␣matrices␣transposed")

set.seed(4000) #Set to number of iterations (n_it=20L,
    setseed(20))
#Running CSD
csd_df <- run_csd(x_1,x_2,n_it=4000L,nThreads=10L,verbose=TRUE)

#Filter result
n_pairs <- nrow(csd_df)
pairs_to_pick <- 1000
index_vector <- seq_len(min(pairs_to_pick,n_pairs))

log_progress("Sorting␣C-values")
c_filter <- order(csd_df$cVal,decreasing = TRUE)[index_vector]
c_frame <- csd_df[c_filter,]

log_progress("Sorting␣S-values")
s_filter <- order(csd_df$sVal,decreasing = TRUE)[index_vector]
s_frame <- csd_df[s_filter,]

log_progress("Sorting␣D-values")
d_filter <- order(csd_df$dVal,decreasing = TRUE)[index_vector]
d_frame <- csd_df[d_filter,]

#CSD filter
cs_filter <- union(c_filter, s_filter)
csd_filter <- union(cs_filter, d_filter)
csd_frame <- csd_df[csd_filter,]

#Writing to files
log_progress("Writing␣to␣file")
write.table(x = c_frame, file = "Her2_hist_4000.txt", sep = ’\t’,
    row.names = FALSE, quote = FALSE)
write.table(x = s_frame, file = "Her2_hist_4000.txt", sep = ’\t’,
    row.names = FALSE, quote = FALSE)
write.table(x = d_frame, file = "Her2_hist_4000.txt", sep = ’\t’,
    row.names = FALSE, quote = FALSE)
write.table(x = csd_frame, file = "Her2_hist_4000.txt", sep =
    ’\t’, row.names = FALSE, quote = FALSE)

#Make histograms for the link scores
#Set the limit for included score values
c_lim = min(c_frame$cVal)
s_lim = min(s_frame$sVal)
d_lim = min(d_frame$dVal)

log_progress("Generating␣histograms")
log_progress("C-link␣score␣historgrams")
jpeg(file="Her2_C_scores_zoom.jpeg")
c = hist(csd_df$cVal,
     main = "Histogram␣of␣C-link␣scores",
     xlab = "C-link␣scores",
     ylab = "Frequency",
     col = "#3A3AD3",
     freq = TRUE,
     breaks = "Sturges")
c$counts[c$counts > 1000] = 1000
plot(c,
     main = "Histogram␣of␣C-link␣scores",
     xlab = "C-link␣scores",
     ylab = "Frequency",
     col = "#3A3AD3",
     freq = TRUE,
     ylim = c(0, 1200))
abline(v = c_lim, col="black", lwd = 2, lty = 5)
dev.off()
plot.new()

jpeg(file="Her2_C_scores_zoom2.jpeg")
c$counts[c$counts > 100] = 100
plot(c,
     main = "Histogram␣of␣C-link␣scores",
     xlab = "C-link␣scores",
     ylab = "Frequency",
     col = "#3A3AD3",
     freq = TRUE,
     ylim = c(0,125))
abline(v = c_lim, col="black", lwd = 2, lty = 5)
dev.off()
plot.new()

log_progress("S-link␣score␣histogram")
jpeg(file="Her2_S_scores_zoom.jpeg")
s = hist(csd_df$sVal,
     main = "Histogram␣of␣S-link␣scores",
     xlab = "S-link␣scores",
     ylab = "Frequency",
     col = "#32CF2E",
     freq = TRUE,
     breaks = "Sturges")
s$counts[s$counts > 1000] = 1000
plot(s,
     main = "Histogram␣of␣S-link␣scores",
     xlab = "S-link␣scores",
     ylab = "Frequency",
     col = "#32CF2E",
     freq = TRUE,
     ylim = c(0,1200))
abline(v = s_lim, col="black", lwd = 2, lty = 5)
dev.off()
```

```
plot.new()

jpeg(file="Her2_S_scores_zoom2.jpeg")
s$counts[s$counts > 100] = 100
plot(s,
     main = "Histogram␣of␣S-link␣scores",
     xlab = "S-link␣scores",
     ylab = "Frequency",
     col = "#32CF2E",
     freq = TRUE,
     ylim = c(0,125))
abline(v = s_lim, col="black", lwd = 2, lty = 5)
dev.off()
plot.new()

log_progress("D-link␣score␣histogram")
plot.new()
jpeg(file="Her2_D_scores_zoom.jpeg")
d = hist(csd_df$dVal,
     main = "Histogram␣of␣D-link␣scores",
     xlab = "D-link␣scores",
     ylab = "Frequency",
     col = "#E02732",
     freq = TRUE,
     breaks = "Sturges")
d$counts[d$counts > 1000] = 1000
plot(d,
     main = "Histogram␣of␣D-link␣scores",
     xlab = "D-link␣scores",
     ylab = "Frequency",
     col = "#E02732",
     freq = TRUE,
     ylim = c(0,1200))
abline(v = d_lim, col="black", lwd = 2, lty = 5)
dev.off()

plot.new()
jpeg(file="Her2_D_scores_zoom2.jpeg")
d$counts[d$counts > 100] = 100
plot(d,
     main = "Histogram␣of␣D-link␣scores",
     xlab = "D-link␣scores",
     ylab = "Frequency",
     col = "#E02732",
     freq = TRUE,
     ylim = c(0,125))
abline(v = d_lim, col="black", lwd = 2, lty = 5)
dev.off()
log_progress("DONE")
```

A.4 Link score details for different iteration numbers

Correlation and variance for each of the link types C, S and D for the two calculations with different bootstrap iterations for each network was performed to asses the stability of the results. The stability comparison was done for the identical gene pairs, excluding those in that did not match with a gene pair in the other calculation. For the BL:HC network these were stable from the start, but as the number of identical gene pairs with a S link type was very low, higher iterations numbers were used, see Table 4.1. The comparison of correlation and variance values done between 2000 iterations and 4000 iterations were stable for the identical gene pairs and was concluded to be stable and sufficient to use in further analysis.

For the remaining networks the same assessment was done and the comparison of identical gene pairs between different bootstrap iteration numbers are listed in Table A.3. For the HNCA:HC network all link types exceeded the threshold of 66% and the assessment of correlation and variance showed stable values for the 2000/4000 bootstrap iteration comparison in the CSD_R calculations. The same was observed for HER2:HC and NL:HC, while LumB:HC and LumB:HC did not have a sufficient number of identical gene pairs of the S link type, when comparing 2000/4000 iterations. A higher number of bootstrap iterations was implemented and LumB:HC exceeded the threshold at the 4000/8000 comparison and LumA:HC at the 8000/16000 comparison. Evaluating the correlation and variance revealed that they were stable in these comparisons and the network generated with 8000 and 16000 iterations was used for analysis of LumB:HC and LumA:HC respectively.

Table A.3: Overview of the gene pairs in the top 1000 selected conserved (C), specific (S), and differentiated (D) link types from the network calculations (HNCA:HC, LumA:HC, LumB:HC, HER2:HC, NL:HC) that, when comparing different bootstrap iteration numbers (B), are identical in both selections (Matches) and the percentage (%). Percentages below 66% are bold.

Network	B	LT	M	%	Network	B	LT	M	%
HNCA:HC	2000/ 4000	C	982	98.2	LumB:HC	2000/ 4000	C	988	98.8
		S	920	92.0			S	593	**59.3**
		D	960	96.0			D	967	96.7
LumA:HC	2000/ 4000	C	987	98.7		4000/ 8000	C	991	99.1
		S	548	**54.8**			S	687	68.7
		D	953	95.3			D	972	97.2
	4000/ 8000	C	989	98.9	HER2:HC	2000/ 4000	C	989	98.9
		S	642	**64.2**			S	893	89.3
		D	960	96.0			D	951	95.1
	8000/ 16000	C	992	99.2	NL:HC	2000/ 4000	C	988	98.8
		S	739	73.9			S	919	91.9
		D	972	97.2			D	958	95.8

A.5 CSD histograms

Histograms of the complete collection of link scores for each of the link types calculated with 20 bootstrap iterations for BL:HC is available in Figure A.2. The corresponding distributions generated with 4000 bootstrap iterations for BL:HC is available in Figure A.3.

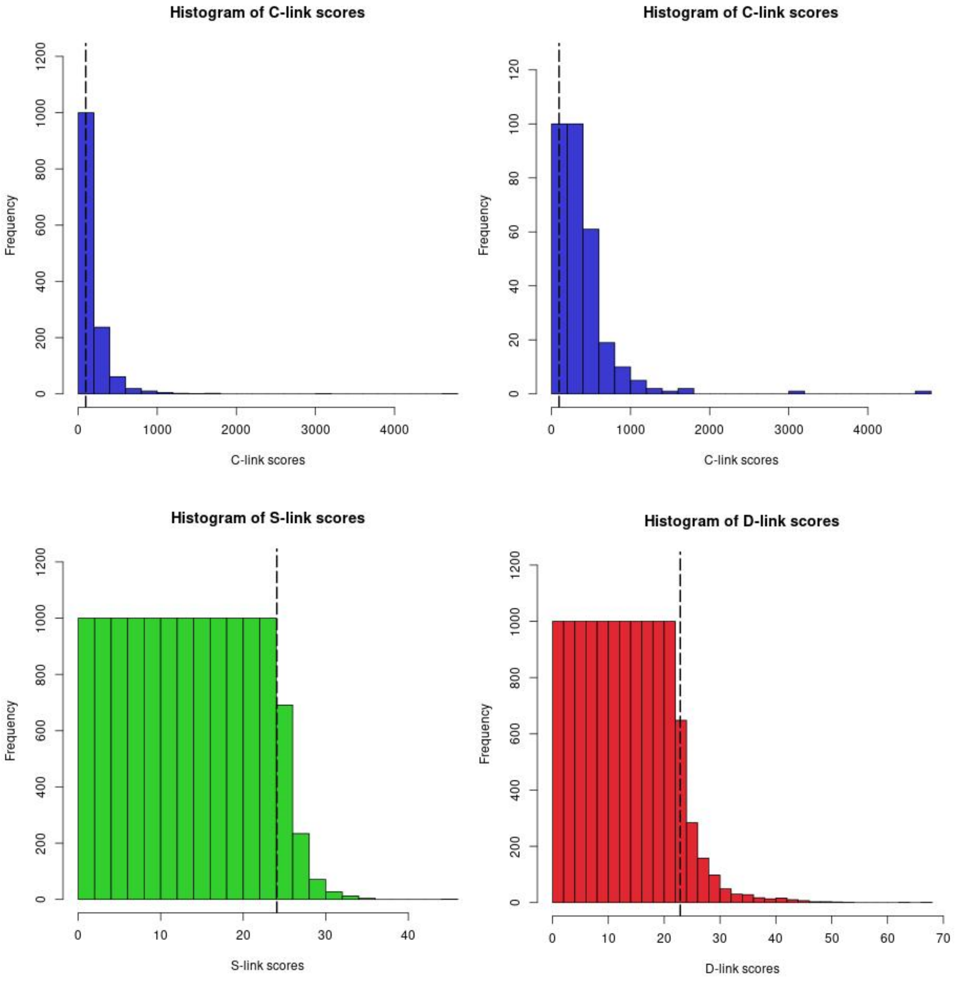


Figure A.2: Histograms of the CSD calculations of BL:HC with 20 bootstrap iterations. From the top left: histogram of C-link scores with a cut-off equal to 1000 for each bar, histogram of C-link scores with a cut-off equal to 100 for each bar, histogram of S-link scores with a cut-off equal to 1000 for each bar, and histogram of D-link scores with a cut-off equal to 1000 for each bar. The dotted lines signifying the score limit of links included in the generated CSD network. The limit for included scores are 95.69 for C scores, 24.07 for S scores and 22.88 for D scores.

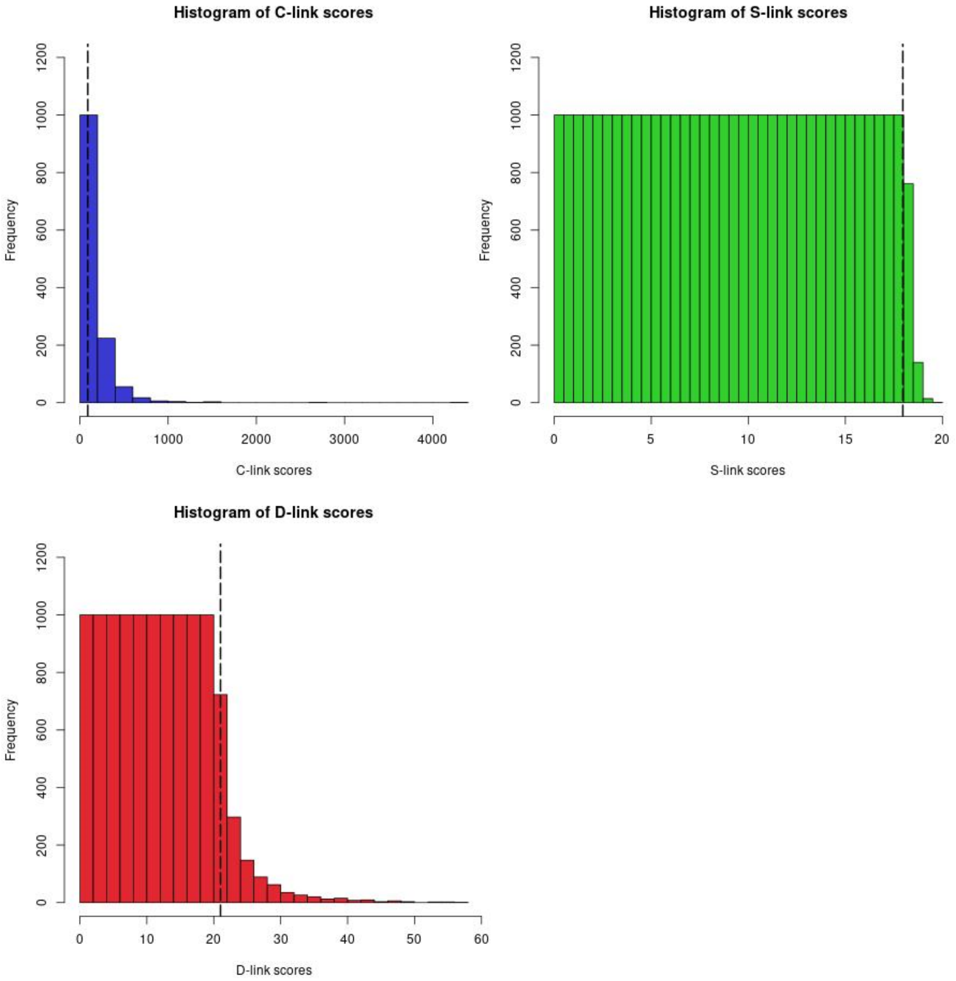


Figure A.3: Histograms of the CSD calculations of BL:HC with 4000 bootstrap iterations. From the top left: histogram of C-link scores with a cut-off equal to 1000 for each bar, histogram of S-link scores with a cut-off equal to 1000 for each bar, and histogram of D-link scores with a cut-off equal to 1000 for each bar. The dotted lines signifying the score limit of links included in the generated CSD network. The limit for included scores are 89.37 for C scores, 17.96 for S scores and 21.00 for D scores.

A.6 CSD Networks of LumA:HC, LumB:HC, HER2:HC and Norm:HC

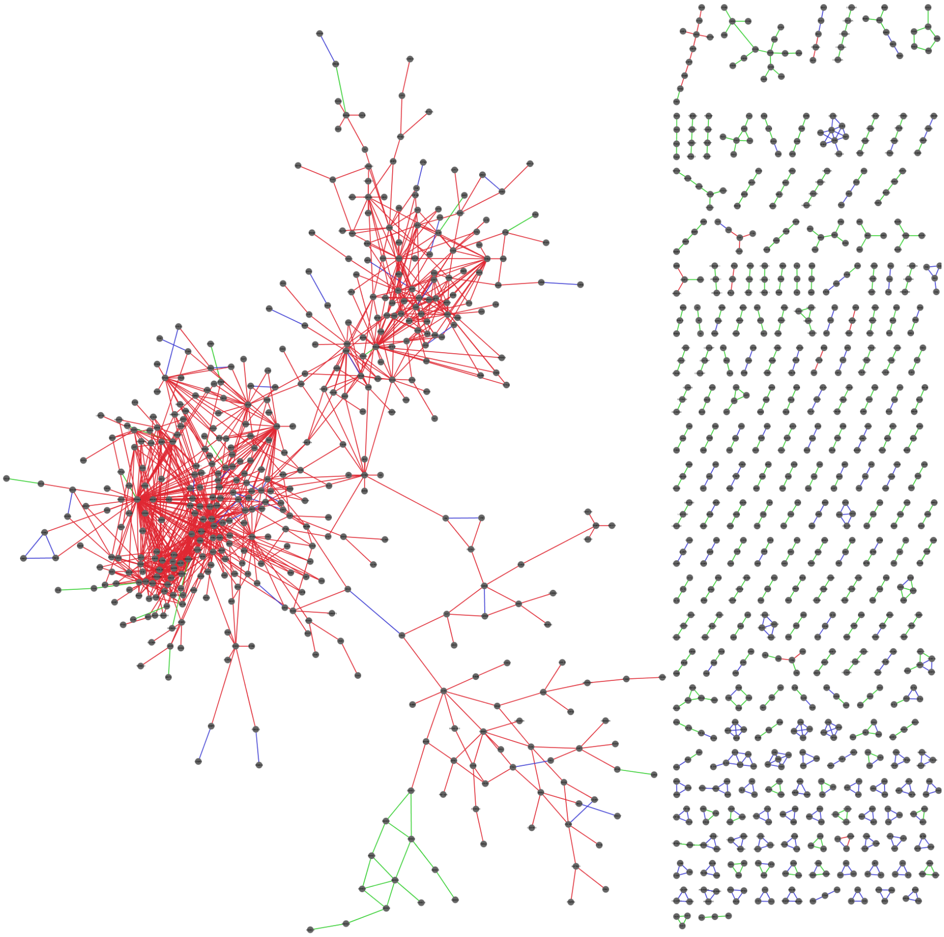


Figure A.4: Visualization of the LumA:HC network with 16000 bootstrap iterations. Links are colored according to their link-type: conserved links are blue, specific links are green and differentiated links are red in correspondence with Figure 2.4. For visual purposes nodes only connected to another node was excluded from this visualization, excluding 2726 nodes and 1363 edges.

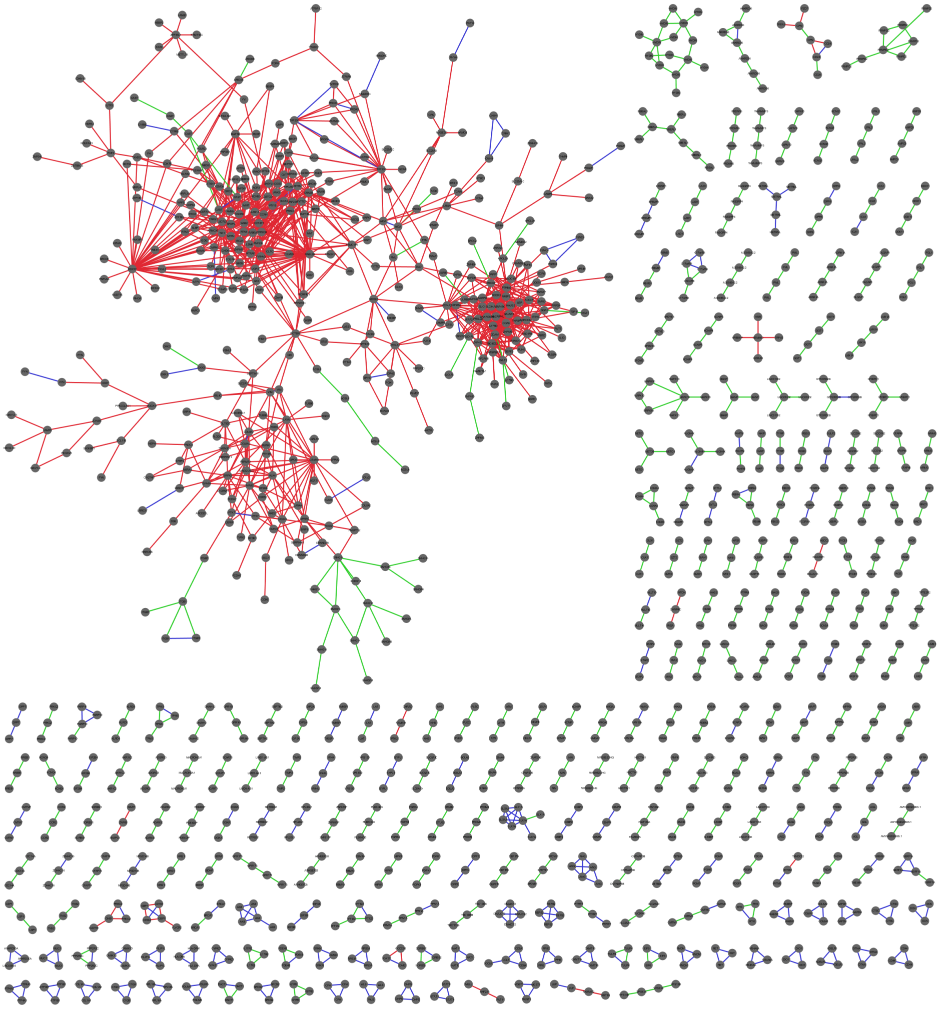


Figure A.5: Visualization of the LumB:HC network with 8000 bootstrap iterations. Links are colored according to their link-type: conserved links are blue, specific links are green and differentiated links are red in correspondence with Figure 2.4. For visual purposes nodes only connected to another node was excluded from this visualization, excluding 2592 nodes and 1296 edges.

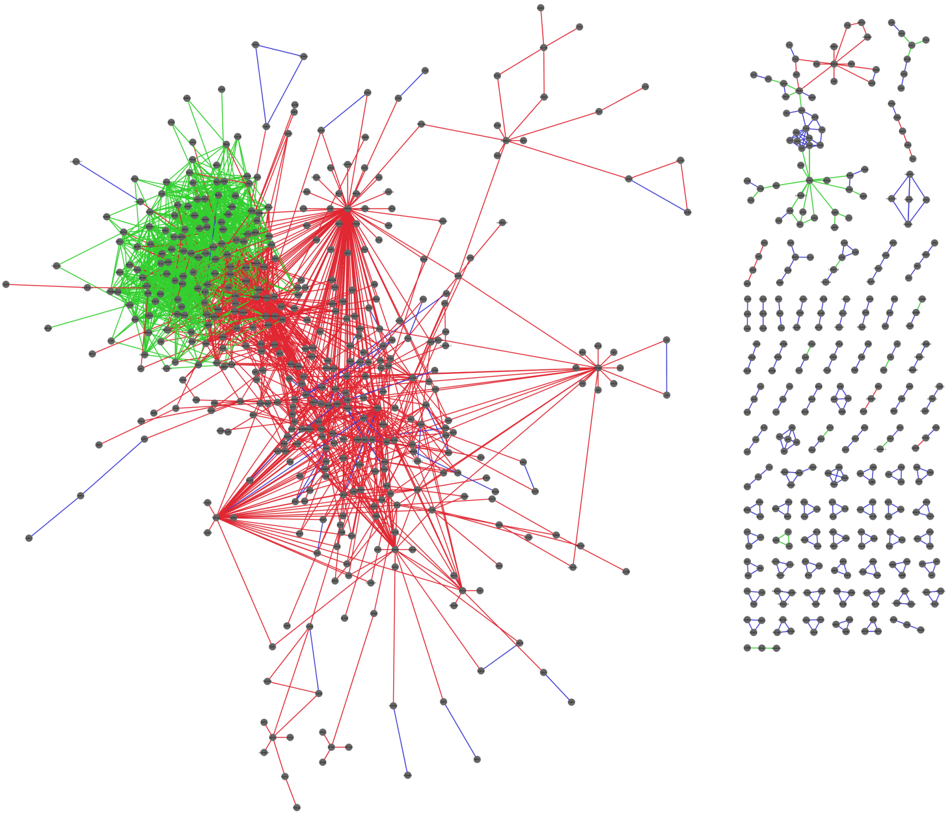


Figure A.6: Visualization of the HER2:HC network with 4000 bootstrap iterations. Links are colored according to their link-type: conserved links are blue, specific links are green and differentiated links are red in correspondence with Figure 2.4. For visual purposes nodes only connected to another node was excluded from this visualization, excluding 1486 nodes and 743 edges.

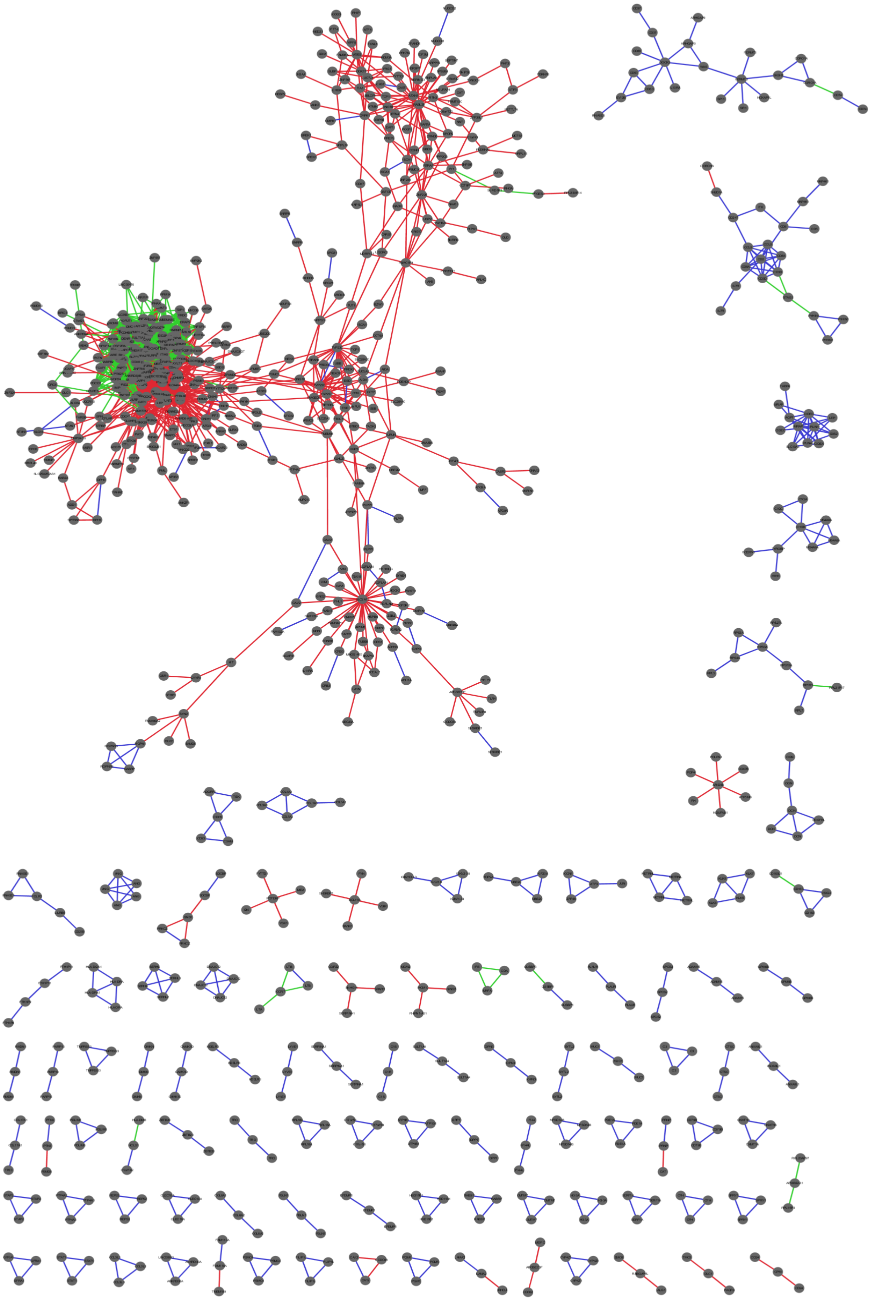


Figure A.7: Visualization of the NL:HC network with 4000 bootstrap iterations. Links are colored according to their link-type: conserved links are blue, specific links are green and differentiated links are red in correspondence with Figure 2.4. For visual purposes nodes only connected to another node was excluded from this visualization, excluding 1272 nodes and 636 edges.

A.7 GO biological processes

General GO biological processes

The general GO biological processes enriched in the CSD networks BL:HC, LumA:HC, LumB:HC, HER2:HC, and NL:HC are given in Table A.4, along with the corresponding fold enrichment (FE).

Table A.4: The top five general GO biological processes enriched in each of the networks with their respective fold enrichment (FE) using the least specific category of biological processes in DAVID. The processes are ordered according to their FE.

Network	General GO category	FE
BL:HC	Biological adhesion	1.6
	Growth	1.5
	Cell killing	1.5
	Locomotion	1.5
	Immune system process	1.4
LumA:HC	Cell aggregation	2.7
	Detoxification	2.1
	Biological adhesion	1.6
	Growth	1.6
	Rhythmic process	1.5
LumB:HC	Detoxification	1.9
	Cell killing	1.6
	Growth	1.6
	Biological adhesion	1.6
	Locomotion	1.5
HER2:HC	Cell killing	2.1
	Detoxification	1.8
	Biological adhesion	1.7
	Growth	1.6
	Immune system process	1.6
NL:HC	Detoxification	2.4
	Cell killing	2.1
	Biological adhesion	1.7
	Locomotion	1.6
	Immune system response	1.6

Specific GO biological processes

The more specific biological processes enriched in the CSD networks BL:HC, LumA:HC, LumB:HC, HER2:HC, and NL:HC are given in Table A.5, along with their corresponding fold enrichment (FE).

Table A.5: The top ten specific GO biological processes enriched in each of the intrinsic breast cancer subtype networks with their respective fold enrichment (FE) from PANTHER. The processes are ordered by their FE.

Network	Biological process	FE
BL:HC	Negative regulation of dendritic cell apoptotic process	7.96
	AMP biosynthetic process	7.96
	Positive regulation of receptor binding	6.68
	T-helper 17 cell differentiation	6.68
	Regulation of protein neddylation	6.08
	T-helper 17 type immune response	6.08
	AMP metabolic process	5.57
	Negative regulation of transforming growth factor beta production	5.57
	Purine nucleoside monophosphate biosynthetic process	4.84
	Purine ribonucleoside monophosphate biosynthetic process	4.77
LumA:HC	Hemidesmosome assembly	8.47
	Negative regulation of translation in response to stress	8.07
	Androgen receptor signaling pathway	5.02
	Regulation of hepatocyte proliferation	4.94
	Cellular response to prostaglandin E stimulus	4,94
	Cellular response to vitamin	4.42
	Cartilage condensation	4.30
	Purine ribonucleoside monophosphate biosynthetic process	4.30
	Negative regulation of blood circulation	4.23
	Negative regulation of intrinsic apoptotic signaling pathway in response to DNA damage	4.14
LumB:HC	Common-partner SMAD protein phosphorylation	9.53
	Optic cup morphogenesis involved in camera-type eye development	8.17
	AMP biosynthetic process	8.17
	Regulation of lamellipodium morphogenesis	7.28
	Regulation of aspartic-type endopeptidase activity involved in amyloid precursor protein catabolic process	6.24
	Regulation of ER to Golgi vesicle-mediated transport	5.34
	Focal adhesion assembly	4.49
	Negative regulation of T cell apoptotic process	4.46
	Response to fluid shear stress	3.93
	Negative regulation of intrinsic apoptotic signaling pathway in response to DNA damage	3.81

Continued on next page

Table A.5 – continued from previous page

Network	Biological process	FE
HER2:HC	Chemokine production	12.58
	Cellular response to interferon-alpha	9.03
	Thymic T cell selection	7.94
	Negative thymic T cell selection	7.62
	Positive thymic T cell selection	7.62
	Negative regulation of transforming growth factor beta production	7.19
	Negative regulation of natural killer cell mediated immunity	7.06
	Negative T cell selection	6.99
	Cell junction disassembly	6.99
	Negative regulation of natural killer cell mediated cytotoxicity	6.52
NL:HC	Negative regulation of dendritic cell apoptotic process	11.97
	Response to L-ascorbic acid	11.17
	Positive regulation of T-helper 2 cell differentiation	9.58
	Regulation of smooth muscle cell chemotaxis	9.58
	Negative regulation of translation in response to stress	9.58
	Negative regulation of transforming growth factor beta production	8.38
	Positive regulation of hepatocyte proliferation	8.38
	Cellular response to prostaglandin E stimulus	8.38
	Hemidesmosome assembly	8.38
	Positive regulation of macrophage cytokine production	8.38

A.8 Network Modules

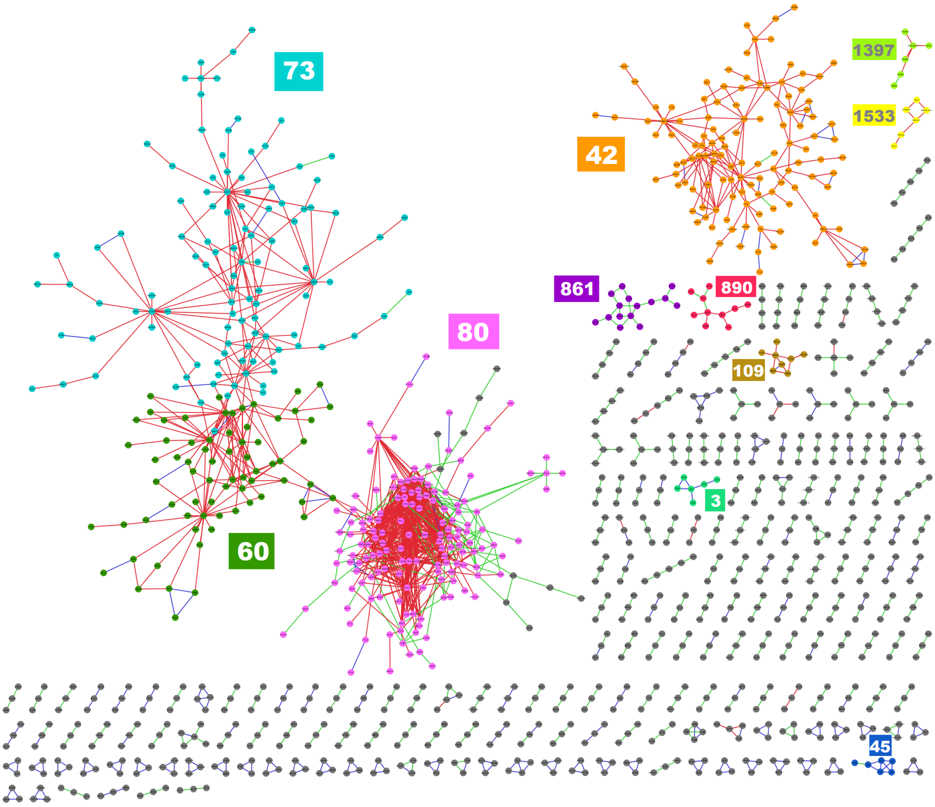


Figure A.8: Communities detected by the Louvain community algorithm [70] in the CSD network BL:HC. Nodes in a community consisting of more than five nodes are colored according to their respective module, and the number denotes the module number. Edges are colored according to their link type, conserved links are blue, differentiated links are red and specific links are green. As in Figure 4.4 genes connected in pairs are excluded for visual purposes.

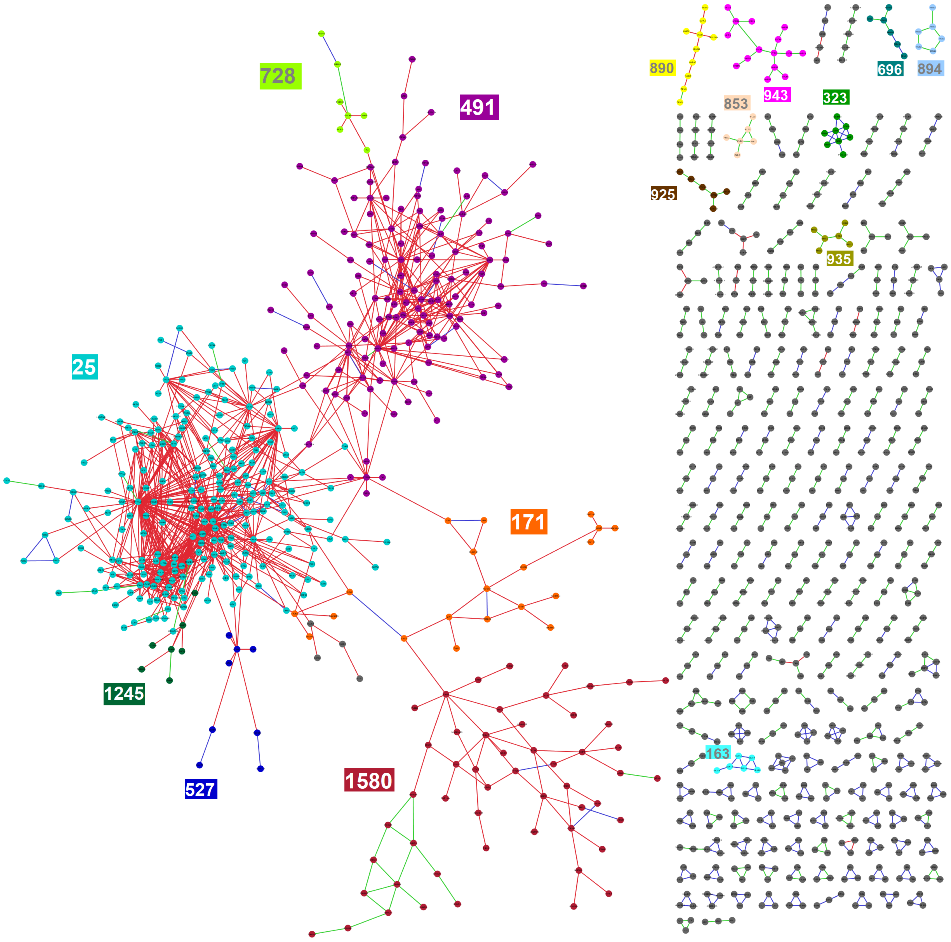


Figure A.9: Communities detected by the Louvain community algorithm [70] in the CSD network LumA:HC. Nodes in a community consisting of more than five nodes are colored according to their respective module, and the number denotes the module number. Edges are colored according to their link type, conserved links are blue, differentiated links are red and specific links are green. As in Figure A.4 genes connected in pairs are excluded for visual purposes.

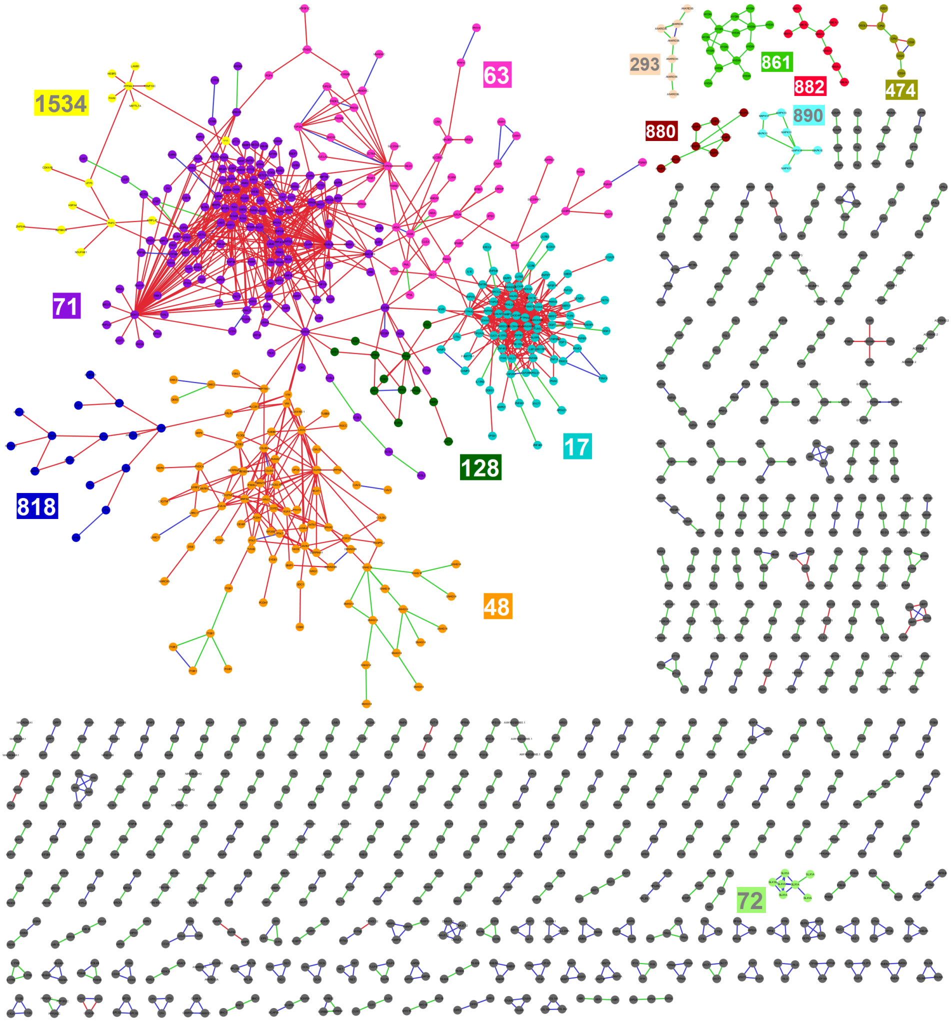


Figure A.10: Communities detected by the Louvain community algorithm [70] in the CSD network LumB:HC. Nodes in a community consisting of more than five nodes are colored according to their respective module, and the number denotes the module number. Edges are colored according to their link type, conserved links are blue, differentiated links are red and specific links are green. As in Figure A.5 genes connected in pairs are excluded for visual purposes.

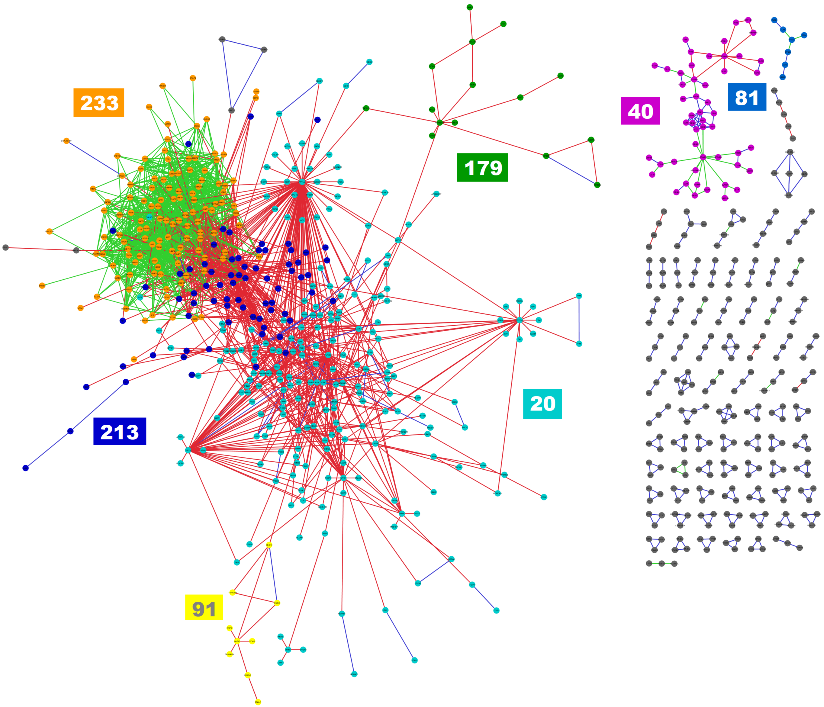


Figure A.11: Communities detected by the Louvain community algorithm [70] in the CSD network Her2:HC. Nodes in a community consisting of more than five nodes are colored according to their respective module, and the number denotes the module number. Edges are colored according to their link type, conserved links are blue, differentiated links are red and specific links are green. As in Figure A.6 genes connected in pairs are excluded for visual purposes.

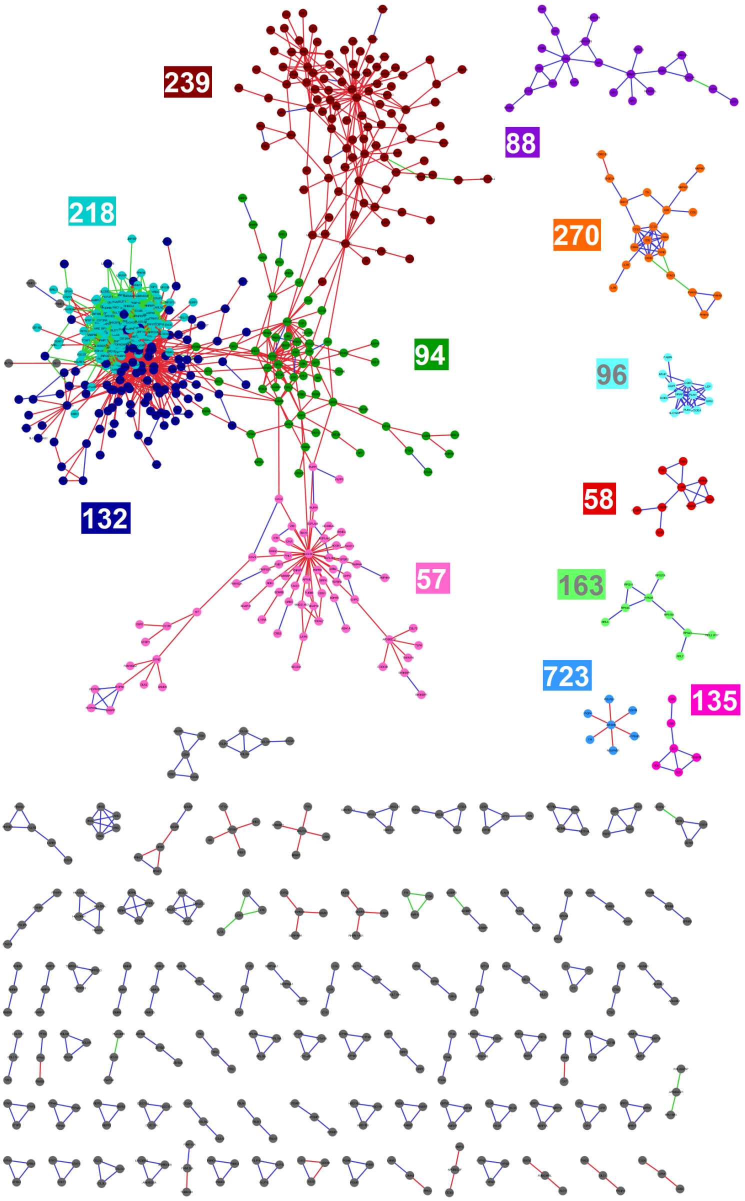


Figure A.12: Communities detected by the Louvain community algorithm [70] in the CSD network NL:HC. Nodes in a community consisting of more than five nodes are colored according to their respective module, and the number denotes the module number. Edges are colored according to their link type, conserved links are blue, differentiated links are red and specific links are green. As in Figure A.7 genes connected in pairs are excluded for visual purposes.

Kristin Salvesen

Using differential co-expression analysis to investigate breast cancer-related tissues

NTNU
Norwegian University of
Science and Technology