Network Based Integration of Proteomic and Transcriptomic Data: Study of BCR and WNT11 Signaling Pathways in Cancer Cells

A dissertation for the award of the degree

Doctor rerum naturalium

of Georg-August University Göttingen
within the doctoral program Environmental Informatics (PEI)
of the Georg-August University School of Science (GAUSS)

submitted by

Maren Sitte
from
Halle (Saale)

Göttingen, 2020

Doctoral Committee:

Professor Dr. Tim Beißbarth
Department of Medical Bioinformatics
University Medical Center Göttingen
Professor Dr. Stephan Waack
Department of Computer Science
University Medical Center Göttingen

Members of the Examination Board:

Professor Dr. Tim Beißbarth
Department of Medical Bioinformatics
University Medical Center Göttingen
Professor Dr. Stephan Waack
Department of Computer Science
University Medical Center Göttingen

Further members of the Examination Board:

1st Referee: Professor Dr. Edgar Wingender
Department of Bioinformatics
University Medical Center Göttingen
2nd Referee: Professor Dr. Burkhard Morgenstern
Department of Bioinformatic
Georg-August University Göttingen
3rd Referee: Professor Dr. Winfried Kurth
Department Ecoinformatics, Biometrics & Forest Growth
Georg-August University Göttingen
4th Referee: Professor Dr. Ulrich Sax
Department of Medical Informatics
University Medical Center Göttingen

Date of oral examination: 30 March 2020

“It is an old saying, abundantly justified, that where sciences meet there growth occurs.“
Frederick Gowland Hopkins

ACKNOWLEDGEMENTS

This dissertation comprises nearly four years of intensive research, and is the result of the effort and support of many people. First of all, I would like to express honest gratitude to my advisor Prof. Tim Beißbarth for his endless support of my PhD study and related research, for his patience, inspiration, and enormous knowledge.

I want to acknowledge my collaborators Prof. Dr. Nils Blüthgen, Prof. Dr. Annalen Bleckmann, Dr. med. Kerstin Menck and Dr. Bertram Klinger for the contributions to the projects and specially for the fruitful discussions. Also, I would like to acknowledge the collaborators, who provided and created the data sets, which are the basis for all analyses in this thesis.

With a special mention to Júlia Perera Bel for all the valuable advices. Many thanks also to Dr. Manuel Nietert. It was fantastic to for all his countless advices and time in which we discussed and debated ideas over a cup of coffee.

Last but not least I thank all my colleags for their constant source of inspiration.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF ABBREVIATIONS

BCR	B cell receptor
BL	Burkitt's lymphoma
BN	Bayesian Network
BIC	Bayesian Information Criterion
DAG	directed acyclic graph
DEA	differential expression analysis
DEG	differentially expressed genes
DLBCL	Diffuse large B cell lymphomas
FDR	False Discovery Rate
JPD	joint probability distribution
KM	Kaplan-Meier
mRNA	messenger RNA
NEM	Nested Effects Model
MFS	metastasis-free survival
OS	overall survival
RPPA	Reverse Phase Protein Arrays
TF	transcription factor

ABSTRACT

Bioinformatics applications in cancer research expanded rapidly over several years in the past. Due to the fast development of high throughput technologies, it became feasible to study the presence of hundreds of genes or proteins measured parallel in one experiment. The challenge is to understand how the regulatory network alters under different conditions or in disease. Their expression values can be used to learn more about their interactions. To study their interplay under different conditions network reconstruction methods were utilized.

This thesis demonstrates a general workflow for integrating data sets from different data sources into a signaling network analysis for cancer cells. Exemplary, BCR signaling in lymphomas and WNT11 signaling in breast cancer was analyzed utilizing gene, proteinn and patient data to elucidate the changes of BCR signaling and WNT11 signaling after specific cell treatment.

The aim of the first study was to investigate proteomic data together with existing gene expression data to predict how lymphomas translate signaling stimuli to expressed phenotypes. BCR-related pathway interplays were reconstructed by analyzing several gene and phospho-protein expression profiles. Therefore, the two network reconstruction techniques NEM and DDEPN were applied to transcriptomic and proteomic measurements, followed by an integrative analysis to identify alterations in BCR signaling after external stimulation.

In the second study, the WNT11 pathways were analyzed in relation to their interplay to one of its receptors ROR2 in human breast cancer. It has been shown that WNT11 signaling highly depends on its receptors and ligands who determine downstream signaling. In an integrative analysis pipeline, transcriptomic and proteo-mic data sets were combined to estimate downstream signaling interplay. Subsequently, patient data was included to associate the findings with clinical outcome.

In both studies, the analysis identified genes, proteins and pathways considered to be biologically important along with potentially new results that can be used to encourage ongoing research.

ZUSAMMENFASSUNG

Die bioinformatischen Anwendungsmöglichkeiten in der Krebsforschung haben sich in der letzten Zeit rasant verbreitet. Durch die schnelle Entwicklung von Hochdurchsatztechnologien wurde es möglich, das Vorkommen von Hunderten von Genen oder Proteinen parallel in einem Experiment zu messen. Die Herausforderung besteht darin, zu verstehen, wie sich das regulatorische Netzwerk unter verschiedenen Bedingungen oder bei Krank-heiten verändert. Die gemessenen Expressionswerte können verwendet werden, um mehr über die Interaktionen zwischen den Genen oder Proteinen zu erfahren. Um ihr Zusammenspiel unter verschiedenen Bedingungen zu studieren, bedient man sich Methoden zur Netzwerkrekonstruktion.

Diese Arbeit zeigt einen allgemeinen Workflow zur Integration von Datensätzen aus verschiedenen Datenquellen in eine Signalnetzwerkanalyse von Krebszellen. Am Beispiel des BCR Signalwegs in Lymphomen und des WNT11 Signalwegs in Brustkrebszellen wurden Gen-, Protein- und Patientendaten analysiert, um die Veränderungen des BCR-Signals und des WNT11-Signals nach einer gezielt durchgeführten Zellbehandlung zu untersuchen.

Das Ziel der ersten Studie war es, Proteomdaten zusammen mit vorhandenen Genexpressionsdaten zu untersuchen, um vorherzusagen, wie Lymphome Signalreize in Phänotypen transformieren. Die Netzwerkinteraktionen zwischen des BCR Signalweges wurden durch die Analyse von Gen- und Phospho-Protein-Expressionsprofile erforscht. Hierzu wurden die beiden Netzwerk-Rekonstruktionstechniken NEM und DDEPN für transkriptomische und proteomische Messungen eingesetzt, gefolgt von einer integrativen Analyse, um Veränderungen im BCR-Signalweg nach externer Stimulation zu identifizieren.

In der zweiten Studie wurden die WNT11-Signalwege in Bezug auf ihr Zusammenspiel mit einem seiner Rezeptoren ROR2 beim menschlichen Brustkrebs analysiert. Es konnte gezeigt werden, dass die WNT11-Signalübertragung stark von seinen Rezeptoren und Liganden abhängt, die die nachgeschaltete Signalweitergabe bestimmen. In einer integrativen Analyse-Pipeline wurden transkriptomische und proteomische Datensätze miteinander kombiniert, um das Zusammenspiel des Downstream-Signalwegs zu unter-suchen. Anschließend wurden Patientendaten einbezogen, um die Befunde mit dem klinischen Ergebnis zu verknüpfen.

In beiden Studien identifizierte die Auswertung Gene, Proteine und Signalwege, die als biologisch wichtig angesehen werden können, sowie potentiell neue Ergebnisse, die zur Weiterentwicklung der laufenden Forschung genutzt werden können.

CHAPTER I

Introduction

1.1 Signaling Pathways

Signal transduction is the process whereby an extracellular stimulus activates a series of signaling molecules inside the cell which finally results in cellular response to the stimulus.
Receptor-mediated signal transduction is an elemental cellular process. The cell membrane acts as a filter to the outside environment and transmits selected stimulatory cues. When a ligand binds to a specific receptor on the cell surface, it alters the shape and activity of the receptor, triggering a change inside of the cell, regulating even changes of gene expression that occur in the nucleus. Or more precisely, the signal is passed down to a special family of proteins, called transcription factors (TFs), which then regulate the expression of genes. Accordingly, TFs are the biological connection between the signaling pathway and the genes. [*Alberts et al.*, 2007]

Errors in signaling interactions are the basis of diseases such as cancer. Therefore, it is becoming increasingly important for future therapies to target disease-specific alterations of cell-signaling mechanisms. [*Wang et al.*, 2013; *Teiten et al.*, 2007]

These intercellular signaling pathways are now among the most studied systems in biology due to their predominant and divergent roles, and their general conservation across species. At the molecular level, a lot of work was invested in identifying their ligands, receptors, intracellular effectors, transcription factors, and modulators. However, the connections between the different signal cascades and their activation are insufficiently understood.

1.1.1 B Cell Receptor Signaling

The diverse B cell receptor (BCR) signaling has been studied extensively. Its complexity leads to many different events, such as cell survival, apoptosis, proliferation, and differentiation into antibody-producing cells or memory B cells. [*Han et al.*, 2003; *Yamanashi et al.*, 1997; *Engel et al.*, 1995] BCR is crucial for normal B cell development and maturation. As the vast majority of B cell lymphomas express the receptor, BCR, and its downstream signaling pathway molecules are attractive therapeutic targets.

The BCR is a connection between membrane immunoglobulin and the IG-Alpha and IG-Beta heterodimer. Anytime an antigen binds to the BCR transmembrane receptor, it initiates the intracellular signaling cascade. Immediately after antigen binding, BCR triggers over phosphorylation of some members of the SRC-family kinases leading to the activation of LYN and SYK. [*Dal Porto et al.*, 2004; *DeFranco*, 1997; *Kurosaki and Kurosaki*, 1997; *Reth and Wienands*, 1997] This activation initiates the regulated aggregation of intracellular signaling molecules. Among them are phosphoinositide 3-kinase (PI3K) and Bruton's tyrosine kinase (BTK). SYK, BTK, and PI3K are crucial members within the BCR signaling cascade that have been investigated as important targets of novel agents. [*Fowler and Davis*, 2013]

Structurally homologous to SYK is ZAP70, which plays a central role in signal transduction from the T cell receptor. [*Chu et al.*, 1998] In B cells, most of the phosphotyrosine activation cascade relies on SYK, but ZAP70 was identified in a subset of normal B cells [*Nolz et al.*, 2005]. More recently, *Crespo et al.* [2006] detected the expression of ZAP70 in some B-cell lines and Burkitt lymphoma.

A result of phosphorylation by SYK and LYN at the Y551 site of the kinase domain, BTK activation is augmented over autophosphorylation of the Y223 site in the SH3 domain. [*Park et al.*, 1996] Additionally, BTK intensifies its activation by engaging the phosphatidylinositol-4-phosphate 5-kinases (PIP5Ks), [*Saito et al.*, 2003] which later ends in persist recruitment of BTK.

Moreover, following BCR ligation the MAPK pathway is activated by various processes. [*Hashimoto et al.*, 1998] The MAPK pathway regulates several transcription factors, including c-Myc through ERK, c-JUN, JNK, p38 MAPK, and MAPK.[*Johnson and Lapadat*, 2002] For instance, in the ERK/MAPK module, activated Raf phosphorylates MEK and the activated MEK subsequently phosphorylates than ERK1/2. [*Dhillon et al.*, 2007] The AKT pathway likewise contributes to BCR-induced survival. AKT is activated when PIP3 is formed by PI3K. By that, AKT gets phosphorylated and organized at the plasma membrane. [*Bellacosa et al.*, 1998] AKT then promotes cell survival by phosphorylating among others the proapoptotic proteins Bad and by intensifying nuclear aggregation of NFAT through inhibition of glycogen synthase kinase 3 (GSK-3). [*Gold et al.*, 1999]

The canonical NF-κB pathway is also an essential contributor to BCR signaling. After stimulation via BTK, PI3K, or AKT, the IκB kinase complex induces phosphorylation of I-κB, promoting nuclear translocation of NF-κB and gene transcription. NF-κB activates a broad collection of genes, which are liable for proliferation, and B-cell survival. [*Balaji et al.*, 2018]

Notably, the diverse relationships between the above described pathways emphasize

the complex structure of BCR signaling. This indicates that there may be many alternatives for possible targets for inhibition.
Another important aspect is the fact that recent biological studies implicate the existence of several feedback regulatory circuits involved in the above mentioned pathways. [*Dougherty et al.*, 2005; *Reth and Brummer*, 2004] As these signaling interplays and feedback mechanisms can block or attenuate treatment efficacy, computational network simulation models can help to better predict alterations by environmental changes including treatment responses.

1.1.2 WNT11 Signaling Pathways

The WNT11 (acronym for wingless-type MMTV integration site) signaling pathway is an evolutionarily highly conserved pathway that orchestrates not only cell fate determination, but also migration, and polarity, among many other functions. [*Komiya and Habas*, 2008] It has an important role during embryogenesis as well as in adult stem cell development and cancer.

Conventionally, WNT11 signals are distinguished by their capability to either stabilize β-catenin in the nucleus (canonical/β-catenin-dependent) or evoke different lines of intracellular signaling independent of β-catenin stabilization (non-canonical). Additionally, the current model is that co-receptors are required for the activation of the different signaling cascades through scaffold proteins such as Disheveled (DVL). [*Komiya and Habas*, 2008; *Kikuchi et al.*, 2009]

In the canonical (β-catenin-dependent) WNT11 pathway, WNT11 signaling inhibits the degradation of β-catenin, which can regulate transcription of a number of genes. The Wnt/β-catenin pathway is initiated by evolutionarily conserved growth factors of the WNT family. Canonical WNTs regulate β-catenin through phosphorylation by the regulation of APC/Axin/GSK-3β - complex, which is also called the destruction complex. In the existence of WNT ligand (On-state), the co-receptor LRP5/6 connects with WNT-bound Frizzled (FZD). This leads to activation of DVL, which in turn releases GSK-3β from APC/Axin. Phosphorylated β-catenin is then translocated into the nucleus via other transcription factors. There it binds to LEF/TCF transcription factors to regulate the function of WNT11 target genes. [*Yang et al.*, 2016]

Non-canonical WNT11 signaling, in contrast, also is activated when a non-canonical WNT ligand (e.g. WNT5a or WNT11) binds to a FZD receptor. The non-canonical pathway is further divided into the Planar Cell Polarity (PCP) and the Wnt/Ca^{2+} pathway. FZD receptors act as the main receptor for WNT ligands and engage other co-receptors to activate certain sub-pathways. In recent years, it is becoming more and more evident that the combination of different WNT ligands and receptors determines which intracellular pathway is activated. [*Grumolato et al.*, 2010; *van Amerongen*, 2012] However, it is still ill understood how different WNT ligands and receptors, through their specific binding, control different signaling pathways.

The co-receptors, like ROR2, aid in the binding of WNT11 proteins to the receptor [*Rosso and Inestrosa*, 2013] and determine the downstream effect, initiating one of the pathways. [*Verkaar and Zaman*, 2010]
In PCP signaling, FZD receptors initiate a cascade of downstream effectors such as the small GTPases Rac1 and RhoA or c-Jun N-terminal kinase (JNK). [*Simons and Mlodzik*, 2008]

In case of the Wnt/Ca^{2+} pathway, non-canonical WNT ligands activate heterotrimeric G proteins, which in turn activate phospholipase C (PLC). This releases Ca^{2+}from intracellular stores. A higher concentration of Ca^{2+} trigger the phosphatase calcineurin (CN), which dephosphorylates NF-AT and leads to its aggregation in the nucleus. This pathway plays a role in controlling cell fate and cell migration. [*De*, 2011]

Previous research showed that irregular expression of certain WNT11 pathway members was associated with various breast cancer subtypes. [*Klemm et al.*, 2011; *Henry et al.*, 2015; *Yang et al.*, 2011] For instance, receptor-tyrosine kinase ROR2 is an orphan receptor, belonging to the Ror family of receptor tyrosine kinases. The protein possesses an extracellular cysteine-rich domain (CRD) that resembles the WNT-binding sites of the Frizzled (FZD) proteins and has been shown to bind Wnt5a. [*Oishi et al.*, 2003; *Sato et al.*, 2010]

To summarize, WNT11 signaling pathways are complex interacting signaling networks and their aberrant regulation is crucial for breast cancer developed. Accordingly, learning how alternative WNT11 receptors such as ROR2 interact with known WNT11 signaling components and what intracellular signaling pathways get initiated will give new insights in the research of drug targets.

1.2 Cancer and Cancer Research

Generally speaking, cancer is guided by (epi-)genetic modifications that allow cells to overproliferate by switching off mechanisms that normally regulate survival and migration. Many of the mutations accumulated in cancer cells influence and deregulate signaling pathways that control cell-cycle, cell growth, division, differentiation, and apoptosis. The development of cancerous cells arises from deregulation of all these coordinated cellular pathways. Changes in the tumor micro-environment are also crucial to cancer development as receptors on the surface of the cells engage intracellular signaling pathways. [*Shaw and Cantley*, 2006; *Sever and Brugge*, 2015; *Yuen et al.*, 2012] Therefore it is essential to find promising drugs that target specific intra- and extracellular signaling components.

This can be achieved by a systematic series of perturbations of cancer cell lines by targeted drugs to model drug response or resistance. The response to perturbation is characterized as a relative change in the expression levels for example of genes and (phospho-)proteins. Drugs that target specific signaling proteins are promising agents in the field of cancer treatment. This approach is under ongoing exploration and could have an impact on how future treatments can ba used. [*Wilson*, 2013; *Lenz and Staudt*, 2010]

Mathematical modeling of the signaling network system is an additional approach to the analysis of therapeutic interventions in silico. They can help to identify patient groups, which could benefit from specific treatment options. Modeling approaches, addressing dynamic functions of intracellular signaling networks, have received increased attention in the last couple of years. [*Klipp and Liebermeister*, 2006; *Janes and Lauffenburger*, 2013; *Azeloglu and Iyengar*, 2015] Network models are able to predict the response of cells to perturbations and will be useful to create combinatorial therapies against cancer.

1.2.1 Lymphoma

Lymphoma, or lymphatic cancer, is cancer that begins in lymphocytes (T cells or B cells). According to the WHO classification [*Swerdlow et al.*, 2008], the two principal types of lymphomas are Hodgkin's lymphomas (HL) and the non-Hodgkin lymphomas (NHL). They involve different types of lymphocyte cells. One of the most common subgroups of NHL in children and adolescents is Diffuse large B cell lymphomas (DLBCL), accounting for 3040% of newly diagnosed non-HL. [*Campo et al.*, 2011; *Hochberg et al.*, 2016] DLBCL is an aggressive (fast-growing) lymphoma that can arise in lymph nodes or also outside of the lymphatic system, such as skin, breast, bone, brain, or basically any organ of the body. Even though immunochemotherapies have significantly improved the general curing prospects of patients with DLBCL, a subset of patients with relapsed still suffer from poor outcomes. In times of huge data sets in both omics profiling and systems biology modeling, there is still little impact of the characterization of the individual tumor genome on the clinical management of DLBCL patients to date. The communication of the cell micro-environment with the tumor cells will be the target of novel therapeutic strategies that have to be investigated.

Another aggressive B-cell lymphoma is Burkitt's lymphoma (BL). BL is an extremely aggressive B-cell non-Hodgkin lymphoma characterized by highly proliferative malignant cells. [*Burkitt*, 1969] BL is uncommon in adults, but $30 - 50\%$ of childhood lymphoma are associated with BL. [*Aldoss et al.*, 2008] In this study, the focus lied on the BCR mediated pathway in Burkitt's lymphoma cell line BL2, because signals who are transmitted after BCR activation are key for the survival of B lymphocytes. [*Gauld et al.*, 2002] When BCR signaling is dysregulated, it leads to tumor development by sustaining the cancer cell population. Along with the upregulation of its various components, the BCR pathway is (highly) active in cancer cells resulting in increased expression of the target genes. Normally, antigen binding to B cell receptors accumulates BCR signaling complexes, which initiate downstream signaling through the phosphorylation and ubiquitylation of cellular proteins.

1.2.2 Breast Cancer

Breast cancer is the leading cancer among women and extremely frequent cause of cancer mortality in most developed countries of the world. *Ferlay et al.* [2019]
Breast cancer is classified into three main immunohistochemical subtypes based on the status of molecular markers for estrogen (ER) or progesterone receptors (PR) and human epidermal growth factor 2 (HER2): hormone receptor positive/HER2 negative, HER2 positive, and triple-negative (tumors lacking all three standard molecular markers).
Recent validation of these molecular phenotypes associates them with treatment respon-se and clinical outcome. [*Perou et al.*, 2000; *Sørlie et al.*, 2001] Triple-negative breast cancer is more likely to recur with local relapse or with distant metastases than the other two subtypes. [*Foulkes et al.*, 2010; *Haffty et al.*, 2006]

Gene expression analysis of various tumor samples via hierarchical clustering has established an alternative subdivision into five (intrinsic) tumor subgroups: basal-like, HER2-enriched, luminal A, luminal B, and normal-breast-like. [*Sørlie*, 2004] Also, this signature is associated with different survival time and response to therapy. [*O'Brien et al.*, 2010]

Nowadays, treatment planning for each patient relies on several factors including tumor morphology and tumor size, expression of ER, PR and HER2 and presence of lymph node metastases. While these factors are used to guide prognosis and therapy, more investigations are necessary to understand the tumor heterogeneity and identify promising targets for cancer treatment.

Different breast cancer subtypes are characterized by different alterations of the WNT11 pathway. While WNT11 signaling plays a central role in various cellular and developmental processes in normal cells, aberrant expression levels of selected WNT11 pathway players were identified to initiate aggressive breast carcinogenesis. [*Koval and Katanaev*, 2018]
In particular RTK-like orphan receptor 2 (ROR2) functions as an alternative receptor or co-receptor for WNT5A and is involved in WNT5A-induced migration of several cell types during cell development. ROR2 is overexpressed in breast cancers and has tumorigenic activity. [*Ford et al.*, 2013] The physical and functional interaction of ROR2 and WNT5A, have been reported in many studies using mice, cultured cells and in vitro systems. [*Henry et al.*, 2015] first studied the role of ROR2 in basal-like breast cancer patients. They showed that ROR2 is expressed in 87% of primary breast cancers and related to shorter survival. Another study from [*Klemm et al.*, 2011] displayed that β-catenin independent WNT11 signaling takes a crucial part in breast cancers which metastasize into the brain.

The aforementioned findings indicate that WNT11 signaling is capable of initiating breast carcinogenesis. Therefore, key WNT ligands and receptors seem promising targets for future drug discovery against breast cancer and the insight into their precise interplay is of high clinical interest.

1.3 Introduction to omics data

1.3.1 The omics landscape

Cellular processes are strongly regulated in multiple layers, resulting in an organized activity of genes and gene products including messenger RNA (mRNA), transcripts and proteins. Each gene instructs the cell how to assemble the pieces for one specic protein. The DNA, that contains the genetic information, lies inside the nucleus. It is transcribed into a mRNA molecule. The mRNA is smaller and more compact than DNA and is capable to move from the nucleus to the ribosomes. After leaving thenucleus, mRNA undergoes some modifications, including removing unneeded sections. Subsequently, it binds to a specific site on a ribosome, where the information is translated into a chain of amino acids to form a protein. The ribosome will translate the mRNA molecule until it reaches a termination sequence, and the protein is released. [*Alberts et al.*, 2007] This sequence of processes (Figure 1.1) are known as the Central Dogma of molecular biology. When a gene has a mutation, the resulting protein is not properly produced, it is because of some mutation in the gene which provides its instructions.

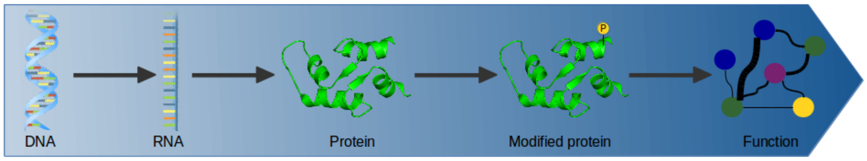


Figure 1.1: *Central dogma of molecular biology.*

Modern biology studies investigate the diverse molecular interactions looking at a wide range of biomolecules and the effects they have. During the last decade, technical improvement of measuring instruments as well as bioinformatics for data analysis has enabled the development of research approaches that intend to discover and quantify large numbers of biomolecules in parallel, drawing a more comprehensive picture of a biological sample's temporal state. This high-throughput analysis of biological samples is commonly referred to as 'omics' and relates to different disciplines in biological sciences, such as genomics, transcriptomics, proteomics, or metabolomics. Each field, or technique, generates plainly distinct information (Figure 1.2) in biological research. In addition, many efforts are made to integrate the different types of data in order to analyze them together.

More precisely, the genome represents the genetic material of an organism, including the coding and the nonconding regions of the DNA. Genomics is the science that studies the structure, function, evolution, and aligning of genomes and addresses the characterization and determination of genes, which direct the production of proteins with the assistance of enzymes and messenger molecules.

The transcriptome defines the exome of a specified cell population. The exome, particularly the protein-coding regions, is defined by the DNA which is transcribed into mRNA. Altogether, these regions, the total exome forms approximately one percent of the human genome. [*Ng et al.*, 2009] Comparing transcriptomes enables the identification of genes that are differentially expressed in different experimental settings.

The proteome is characterized by all expressed proteins under specified conditions. Proteomics is the science that studies proteins in relation to their biochemical properties and functional activities, and how their quantities, modifications, and structures change throughout growth as well as in response to internal and external stimuli.

And finally, the metabolome, which is the terminal downstream product of the genome and is defined as the overall analysis of metabolites in a biological sample. Metabolites are small biomolecules that participate in general metabolic reactions. They are required for the perpetuation and normal function of a cell. [*Goodacre et al.*, 2004]

Altogether, the omics chain with genomics, transcriptomics, proteomics, and metabolomics (Figure 1.2) consists of complex data sets that as an entity comprehensively describe the reaction of biological systems to diseases, genetic variances, and environmental perturbations.

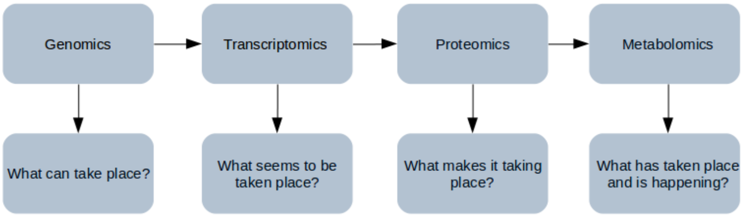

Genomics
Transcriptomics
Proteomics
Metabolomics
What can take place?
What seems to be taken place?
What makes it taking place?
What has taken place and is happening?

Figure 1.2: *The omics chain and its specific research questions. Adapted from [Dettmer et al., 2007].*

One of the advantages of high-throughput data is likewise a major drawback. The large number of measured features might lead to significant findings just by chance. To bypass the difficulties presented by high dimensionality, the data can be grouped in biologically meaningful clusters. [*Tukey*, 1977] This clustering can be achieved by using complementary, but nevertheless methodologically independent, dimension-reduction methodologies or by integrating biological prior knowledge.

Despite many considerable advances in experimental methodologies, data emerging from individual omics approaches are often insufficient to understand gene interactions and functions. Integrated analysis of high-throughput data has been understood as a possible method that can overcome the restraints of individual omics methods and helps in furthering our knowledge of biological systems in their entirety. [*Joyce and Palsson*, 2006]

1.3.2 Omics Data Integration

The concept of 'omics' is now very commonly used in life sciences research. In recent years, the potential to study cellular and molecular systems has been revolutionized as a result of the expansion of omics sciences. For instance, in 2012, the NCIs *The Cancer Genome Atlas* (TCGA) integrated different data types and were able to determine altered modules in three distinct pathways that influence the development of glioblastoma multiform. [*Ciriello et al.*, 2012] These candidate driver mutations can be target to develop new therapeutic options. This study shows the benefit of data integration as these oncogenic alterations were not discovered from data in isolation (either from mutations, copy number changes, or other measurements).

In the same year, R. Chen and his colleagues also demonstrated the benefits of combining different omics data sets in the context of risk detection of type 2 diabetes. In this integrative analysis, the data revealed an increased insulin biosynthetic pathway that spiked during states of viral infections. Their study indicates that viral inflammation can trigger aberrant glucose metabolism and can, therefore, increase the risk of type 2 diabetes. [*Chen et al.*, 2012] Within the scope of their research, they investigated how analysis of the genome, epigenome, transcriptome, proteome, and metabolome can collectively provide advantageous information.

In the field of integrating genomic and proteomic data, there are two general assumptions. The majority of studies, in which both, genome and proteome measurements, are combined, assume that there is an one-to-one relationship between transcript and protein expression.
In an earlier project, Schwanhäusser and colleagues [*Schwanhäusser et al.*, 2011] have looked at RNA and protein separately. To achieve a more accurate observation, they labeled proteins and RNA in mouse fibroblasts. With quantitative mass spectrometry and RNA sequencing, they could calculate absolute mRNA and protein copy numbers in the same samples. Their results suggest that mRNA levels can explain approximately 40% of protein level variation.
In the same direction, more recent studies could demonstrate that there is just a low correlation between transcript levels and protein expression [*Haider and Pal*, 2013; *Zhang et al.*, 2014].

A second inherent assumption is that genome-scale technologies such as next generation sequencing-based transcriptomics and mass spectrometry-based proteomics have equal sensitivity to capture the expression of the gene products. [*Schwartz et al.*, 2018]

Ghosh et al. [2011] performed a study in which they analyzed time-course transcriptome and proteome measurements in order to identify subgroups that respond or not to current anti-HER2 therapy. On the basis of two different omics data sets, they classified distinctive transcriptional and signaling profiles for four patient subgroups associated with response to trastuzumab. They showed that breast cancers driven primarily by HER2 homodimerization are very sensitive to trastuzumab therapy. Consequently, the inhibition of HER2 heterodimerization can increase clinical outcomes (i.e. reduce treatment resistance and risk of disease relapse) in this particular subgroup.

Taken together these examples illustrate the potential of integrating diverse 'omics' data and how it can help the research in biology and medicine. Different methods that aim to integrate heterogeneous data sources have been developed in the last years. The particular methods of focus in this thesis are introduced in section 2.2.3.

1.4 Measurement techniques for transcriptomics and (phospho-)proteomics

The measurement tools comprise of a number of different high-throughput technologies, including DNA microarrays, protein arrays, deep sequencing and mass spectrometry. They allow system-wide unbiased molecular measurements, which can then be used for drug discovery, target validation and the identification of genes or to reproduce the events in an signaling response. This section provides an overview of some of the common measurment techniques within the fields of transcriptomic and proteomic.

1.4.1 Transcriptomic

The analysis of mRNAs provides direct observations of cell- and tissue-specific gene expression characteristics. This information is necessary to gain a better understanding of the dynamics of cellular and tissue metabolism, and to apprehend whether and how adjustments in the transcriptome profiles influence health and disease.
The first effort to study the whole transcriptome began in the beginning of the 1990s. [*Adams et al.*, 1991] Nowadays, the two main gene expression profiling technologies are microarrays and deep sequencing of RNA (RNA-seq) allow the (reproducible) quantification of the abundance of mRNA.

1.4.1.1 Microarrays

The basic principle of DNA microarrays builds on the principle that complementary sequences will bind to each other. Typically, they comprise genomic DNA fragments that are complementary to transcripts of interest. The DNA molecules are labeled with fluorescent markers, which then react with probes of the DNA chip. Next, the target DNA fragments ahead with complementary string attach to the DNA probes. When the remaining DNA fragments are washed away, target DNA sections can be identified by their fluorescence emission captured by a laser beam. The fluorescence intensity at each location on the array indicates the transcript abundance for that specific sequence [*Barbulovic-Nad et al.*, 2006].

Different technologies of DNA microarrays are produced using individual fabrication methods. A frequently used technique to gather transcriptomic data is cDNA microarrays as introduced by [*Schena et al.*, 1995]. They are using polymerase chain reaction (PCR) in the first step and a robot-controlled printer in the second step. Some other similar methods utilize ink-jet like printers to spray chemically synthesized oligonucleotide probes on the microarrays.

Another concept is to synthesize the probes directly on the surface of an array using photo-activated chemistry. Affymetrix GeneChip$^{\text{TM}}$ is one of the most popular microarray chips using this technique. It measures a single sample on one slide and consists of thousands of short oligonucleotide probes spotted on a solid substrate. The arrays consist of a highly ordered matrix of hundreds of thousands of oligonucleotides. They contain more than 33000 genes with over one million oligonucleotides. The approach leans on light-deprotection of the growing oligonucleotide. In each step, the oligonucleotides are built one base after the other. The individual sites on the array bind to the next nucleotide (A, T, C or G) and are marked using photo-activated chemistry. One data set analyzed in this work is based on this technology (see section 2.5.1.1).

1.4.1.2 RNA-Sequencing

Studies utilizing RNA-Sequencing have already transformed our view of the amount and complexity of transcriptomes. Contrary to microarrays, RNA-seq is not restricted to the hybridized probes. Using deep-sequencing technologies, it allows measuring genome-wide expression levels, independent of annotated regions.

In general, a library of cDNA is constructed from a sample's RNA with adaptor molecules attached to one or both ends. Each molecule is then sequenced to gather short sequences from one end (single-end) or both ends (pair-end) by sequential hybridization readout. The sequencing performs successive cycles of base incorporation, washing, and imaging. The lengths of a readout is usually between 50 and 700 bp. In the subsequent bioinformatic pipeline, the reads are quality checked and aligned to a reference genome or transcripts. Alternatively, if a reference is not yet available, the reads are assembled *de novo* without a genomic sequence.

Compared to microarrays, input RNA quantity is much lower for RNA-seq, which allows better investigation of cellular structures, down to the single-cell level when combined with linear amplification of cDNA. [*Hashimshony et al.*, 2012] Furthermore, in contrast to microarrays, RNA-seq is not limited to the hybridized probes but allows to measure genome-wide expression levels, independent of annotated regions. A third advantage is the possibility to detect isoforms. [*Malone and Oliver*, 2011]

The RNA-seq technology was used within the study of WNT11 signaling in breast cancer (see section 2.5.2.2.1).

1.4.2 (Phospho-)proteomic

The direct analysis of protein expression was commonly accomplished on a small scale, using, for example, immuno- or two-hybrid assays. That limited the analysis volume to just a few proteins. The latest measurement technologies have absolutely pushed our knowledge of biochemistry and cell biology, including protein dynamics, multiprotein complexes forward [*Picotti and Aebersold*, 2012] and found application in cell signaling research [*Collins et al.*, 2007].

Phosphorylation is one of the main mechanisms of post-translational regulation of proteins and a large percentage of proteins are phosphorylated at some stage during their life cycle. Phosphorylation causes the protein to become activated (or deactivated) and enables it in turn to initiate the phosphorylation of other proteins in the cascade, finally causing cell-wide changes such as apoptosis, cell differentiation, and growth.
Accordingly, studying the protein post-translational modifications, especially phosphorylation, empowers the discovery of signaling network alterations guided by genomic changes. As a result, quantitative measurement of changes of phospho-protein plays a growing role in studying signaling pathways in a cell. It also improves our understanding of cellular responses to external and internal stimuli.

Over the last years, proteomics has experienced a huge development in methodologies in the direction of large-scale study. Also, the field of phosphoproteomics has developed to larger scale approaches. There are two main measurement methods in phosphoproteomics: antibody- and mass spectrometry-based. A broad summary of these methods can be found in [*Terfve and Saez-Rodriguez*, 2012]. In brief, antibody-based methods are generally specific and depend on the quality of the antibody. They are suitable to measure time courses of target proteins across many conditions. [*Lee et al.*, 2012] To date, most commonly used antibody-based technologies are protein arrays, reverse-phase protein arrays, and the bead based xMAP technology from Luminex. [*Saez-Rodriguez et al.*, 2011] However, the number of targets that can be measured is limited. In comparison, mass spectrometry techniques enable the systematic identification and quantification of phosphorylated proteins. On the downside, they require expensive equipment and expert knowledge for the often elaborate protocols. [*Steen et al.*, 2006] In the following, an introduction to the two aforementioned antibody-based methods is given in more detail.

1.4.2.1 Luminex Bio-Plex® Assays

The Bio-Plex assays use the Luminex xMAP® technology, which means that they use an antibody sandwich for detection. Immunoassays based on Luminex xMAP is a high-throughput technology, which allows simultaneous quantification of multiple secreted proteins. The Bio-Plex system used here is based on the principles of fluorescence imaging. This technique consists of 3 main steps.

In step one color-coded beads, labeled with analyte-specific capture antibody for the protein of interest, are added to the assay. In a next step, the antibodies capture the analyte of interest. Then biotinylated detection antibodies specific to the analyte are added and compose an antibody-antigen sandwich. Additional Phycoerythrin (PE)-conjugated Streptavidin (SE) is added as a signal for the measurements.
In the last step, beads are read with dual-laser flow-based detection. One laser classifies the bead and determines the analyte based on bead color and the second laser quantifies the signal through measuring the reporter molecule PE-SE. This signal is proportional to the amount of bound analyte. [*Bio-Plex*, 1999; *Houser*, 2012] The investigated Luminex Bio-Plex data set within this thesis is described in section 2.5.1.2.

1.4.2.2 Reverse Phase Protein Arrays

An established technique for the simultaneous analysis of different proteins is Reverse Phase Protein Arrays (RPPA). RPPA measures levels of protein expression, as well as protein modifications such as phosphorylation and therefore allows studying the activation status of cell signaling pathways. It already has been used quantitative analysis of protein expression in cancer cells, cell signaling analysis and clinical prognosis or therapeutic prediction. [*Nishizuka et al.*, 2003; *Spurrier et al.*, 2008; *Ummanni et al.*, 2014]

RPPA was introduced by *Paweletz et al.* [2001] as a reproducible technology. Usually, with microarrays, the samples are directly spotted on the slides. In contrast, the RPPA technology is a type of protein microarray that comprises a reverse method.
The biological samples of interest are lysed, producing a homogeneous mixture (lysate), and these lysates are printed onto an array according to a dilution series. These arrays are typically glass plates. On one side they have a nitrocellulose membrane and the lysates are printed on the nitrocellulose. In order to measure the protein of interest, the array is first interrogated with an antibody specific to the protein of interest (the primary antibody). After binding is completed, loose material is washed away. In the second step of incubation, the array is interrogated with a fluorescently labeled antibody (a secondary antibody) which recognizes the primary antibody. Afterward, the slides are scanned and a microarray image analysis software is performed. By comparing the relative level of fluorescence, differential protein expression across all the samples on one slide can then be evaluated simultaneously.

Even though RPPA data is restricted to the selected antibodies and profiles a smaller, predefined set of proteins, antibody microarrays are currently seen as a worthwhile method in view of their small required quantities, affordability, multiplexed detection power, rapidness, and automatization. [*Alvarez-Chaver et al.*, 2014] The details for the RPPA data set used in this thesis are specified in section 2.5.2.2.2

1.5 Statistical methods for the analysis of omics data

Statistical analysis can help to extract information, that is not directly observable. Various models are available of which some will be described in this section. Methods from statistics, including differential expression analysis, and machine learning, such as clustering, are more ”descriptive” approaches in the sense that they help to characterize the data. ”Predictive” concepts are used to estimate the behavior of a system under specified conditions.Methods for this are regression approaches such as linear and logistic regression. More complex models can be built that include mechanistic or causal relationships between members of the system, that can be described by a graph (”network diagram”). Such models involve differential equations, logic-based, and Bayesian network models. Here, the methods relevant for this thesis are introduced.

1.5.1 Differential Expression Analysis

Differential expression analysis (DEA) consists of two main tasks: First, estimate the magnitude of differential expression between two or more conditions based on expression levels from replicated samples, that means, calculate the (logarithmic) fold change. Secondly, estimate the significance of the difference and correct for multiple testing.

The methods were originally developed for microarray data, e.g., in limma. Limma is a R package for DEA of data collected from microarray experiments. The main concept is to fit a linear model to the expression data for each gene or protein. The method uses empirical Bayes to obtain information across genes or proteins to make the analyses more robust for experiments with a just small number of arrays. [*Smyth*, 2004]

There are different methods for RNA-seq data, (such as edgeR [*Robinson et al.*, 2010] and DESeq/DESeq2 [*Anders and Huber*, 2010; *Love et al.*, 2014]) based on negative binomial (NB) distributions or (baySeq [*Hardcastle and Kelly*, 2010] and EBSeq [*Leng et al.*, 2013]) which are Bayesian approaches based on a negative binomial model.
The best performing tools tend to be edgeR, DESeq/DESeq2, and limma-voom [*Ritchie et al.*, 2015] (for reviews of DGE tools see [*Rapaport et al.*, 2013; *Soneson and Delorenzi*, 2013; *Schurch et al.*, 2016]). DESeq and limma-voom turn to be more conservative than edgeR, because they better control of false positives. edgeR is recommended for experiments with fewer than 12 replicates [*Schurch et al.*, 2016].
These tools are implemented in the R language and realize various statistical methods that have been developed during the past decades. The underlying approach in each of them is the same: the gene expression difference for a given gene is estimated using regression-based models. The statistical tests assume the null hypothesis of no effect is true. In other words, it is tested against the hypothesis that the difference is close to zero which means that there is no difference in the gene expression values that are not observed randomly.

High-throughput data sets have usually many more features (genes) than cases (patients or experiments), which results in a high risk of overfitting. To avoid overfitting one might control the False Discovery Rate (FDR). FDR is defined as the expected value of the proportion of false positive features among all of those significant features Benjamini and Hochberg introduced the idea of a FDR to control for multiple hypothesis testing. Controlling FDR increases the power of the method. [*Mathur et al.*, 2011; *Benjamini and Hochberg*, 1995]

1.5.2 Network Analysis

Biological processes can be modeled as a network of causal influences utilizing information from different sources. Mathematical and computational methods are required to organize the overwhelming quantity of data and to make interpretable. Network reconstructions are effective strategies, to obtain a comprehensive interpretation of the results of differential expression analysis. A lot of effort has been invested into learning networks and pathways from gene or protein expression data and prior knowledge. In this section, common network analysis approaches are introduced. Although, a general introduction to networks, with special focus on Bayesian networks, is given. In section 2.2, some specialized types of Bayesian networks that are relevant for this thesis are provided.

1.5.2.1 Network Analysis Methods

In bioinformatics, network methods, have been used to study gene expression data [*Friedman et al.*, 2000; *Yu et al.*, 2004], predict protein-protein interactions [*Jansen et al.*, 2003], infer protein signaling networks [*Friedman*, 2004; *Sachs et al.*, 2005; *Bradford et al.*, 2006], cancer recurrence [*Rouprêt et al.*, 2008] and to infer the statistical dependency between perturbation experiments [*Maathuis et al.*, 2009]. Network analysis consists of various deterministic and probabilistic methods to infer regulatory dependencies from experiments with interferences in the cellular processes.

One common approach is Boolean networks. S. Kaufman [*Kauffman*, 1969] firstly introduced Boolean networks for qualitative description of gene regulatory interactions. Since then Boolean networks have become a versatile research field. They are directed graph, where each node represents a gene and can be either 0 or 1. A Boolean function models the parent states to its child state. Perturbation on distinct regulators allows to infer the architecture of Boolean networks. [*Ideker et al.*, 2000]

Network Component Analysis (NCA) is a network structure-driven framework for inferring regulatory signal dynamics. Unlike classic statistical concepts like independent component analysis or principal component analysis, NCA employs the (connectivity) structure from transcriptional regulatory networks to restrict the decomposition to a unique solution. [*Liao et al.*, 2003; *Tran et al.*, 2005]

Correlation-based graphs assume that the correlation analyses reflect a coordinated interaction between genes (vertices) across the data set. [*Rice et al.*, 2005; *Batushansky et al.*, 2016] Partial correlation coefficients have also been used to identify novel gene networks through the minimization of redundant edges in the network. [*de la Fuente et al.*, 2004; *Veiga et al.*, 2007]

Rather than correlating one relation with another, one may want to predict one relation knowing the other. A way to answer this question is regression [*Segal et al.*, 2003; *Huynh-Thu et al.*, 2010] and shrinkage techniques [*van Someren et al.*, 2006]. However, their weakness can be observed when the number of variables is large. Then, they mix direct and indirect associations. [*Zuo et al.*, 2014] For instance, a strong correlation for gene A with B and A with C will predict a less strong but probably still statistically significant correlation for gene pair B and C. As a consequence, when the number of genes increases, these networks are likely to over-estimate the network with too many false positives.

Another widely used approach to model gene regulatory network are Bayesian Network (BN) models. BNs and variations are today the focus of research that deals with discovering novel interactions, information dependencies and regulatory relationships from expression data. The advantage of using BNs is that by modeling conditional dependence relationships, BNs only identify direct associations. Nevertheless, learning the structure of Bayesian networks for data of high dimensions takes time and can be statistically inaccurate. Additionally, BNs cannot model cyclic structures, such as feedback loops, which occur frequently in biological networks. [*Friedman et al.*, 2000] In section 2.2.0.1 a more detailed description of this method, is provided.

In some biological frameworks, resulting measurements fail to precisely reconstruct the underlying network. In such situations, it is beneficial to integrate prior knowledge coming from literature about gene or protein interactions into the network model into network reconstruction. [*Werhli and Husmeier*, 2008; *Bender et al.*, 2011; *Eduati et al.*, 2012; *McDermott et al.*, 2013] Such restraints cut down the computational costs and assure that approved interactions are considered in the final model.

1.5.2.2 Using Network Databases as prior Biological Knowledge in Network Reconstruction

In general, reconstructing networks from expression data is a challenging question that has become crucial for the understanding of complex regulatory processes in cells. In addition to data-driven network models, there is a growing number of databases [*Bader et al.*, 2006] that capture pathway information in high detail. From publicly available databases such as STRING [*Franceschini et al.*, 2013], KEGG [*Kanehisa and Goto*, 2000], BioGRID [*Stark et al.*, 2006], and ConsensusPathDB [*Kamburov et al.*, 2011], one can obtain numerous types of interactions including protein-protein, signaling, and gene regulatory interactions. Biological networks reconstructed from these databases were found to be valuable. For instance, *Chuang et al.* [2007] reconstructed protein-protein interaction (PPI) network from multiple databases to help identify markers of metastasis for breast cancer studies using gene expression data.

In high-throughput experiments, each sample is described by the expression levels of thousands of genes, or proteins. The large amount of variables not only gives a great opportunity to identify a broad range of biological processes, but also, rises serious (statistical) challenges. Generally, classic statistical methods estimate connections between variables based on mathematical criteria, such as correlation. By that, they cannot differentiate between correlation that comes from a biological source and random correlation caused by the high-dimensionality of the data and measurement noise. Furthermore, variations in expression values can also arise from a biological variation of the studied object. Therefore, a challenge in analyzing high-throughput data is to consider the different variation sources. [*Reshetova et al.*, 2014]

Recent approaches [*Ghanbari et al.*, 2015; *Li and Jackson*, 2015; *Stavrakas et al.*, 2015; *von der Heyde et al.*, 2016] apply prior biological knowledge. The intention of these methods is to guide the statistical analysis to decrease the detection of spurious relations. Additionally, prior knowledge may be used to test the compatibility of experimental data and existing knowledge to compensate for potential gaps or include extra information. The links between variables (genes or proteins) can be resolved, for instance, from the aforementioned databases.

1.5.2.3 Visualization of Gene and Protein Networks

Since the graphical representation of gene and protein networks may highlight important substructures, visualization is more and more used to study the underlying graph structure of the biological networks, such as phylogenetic trees, protein-protein interaction networks, metabolic networks or genetic regulatory networks. [*Junker and Schreiber*, 2008]

Given a specific graph, modern layouts algorithms are optimized for speed and aesthetics. In particular, they seek to minimize overlaps and edge crossing, and design symmetric substructures to facilitate the reading of a graph. Such algorithms are e.g. layered graph drawing methods, also known as Sugiyama-Tagawa-Toda algorithm [*Sugiyama et al.*, 1981], which positions nodes on the levels of a hierarchical layout and the group of algorithms based on the force-directed layout [*Fruchterman and Reingold*, 1991]. In circular layout methods [*Doğrusöz et al.*, 1997], the vertices of the graph get arranged on the circumference of a circle in a way that reduces edge crossings.

In the last years, many software tools for network visualization were developed. Three of the most common tools are:

(i) *Cytoscape* [*Shannon et al.*, 2003] is a software platform to visualize molecular interaction networks and allows to integrate for example gene expression profiles.

(ii) *NetworkX* [*Hagberg et al.*, 2008] is a Python package, which allows studying the structure, dynamics, and functions of networks.

(iii) There are multiple packages implemented in the functional programming language R. The *statnet* set of packages [*Handcock et al.*, 2003] provides functions for the analysis of a wide range of network data coming from diverse areas. Another popular R package is *igraph* [*Csardi and Nepusz*, 2006] which is a library collection for creating and manipulating graphs and analyzing networks. It is also available as Python package. A third R package is called *Rgraphviz* [*Hansen et al.*, 2019]. It provides a connection between R and the third-party software *graphviz* [*Ellson et al.*, 2002].

1.6 Aims and Motivation

Driven by the observations that cellular processes constantly result in multiple and complex responses [*Westerhoff and Palsson*, 2004], and catalyzed by the flood of omics data that were accessible, systems biology emerged in recent years. Systems biology connects experimental, theoretical, and modeling techniques to study biological organisms at all levels, from the molecular to the cellular level. [*Kitano*, 2002] It is applied in a wide variety of fields from plant biology over inflammatory disease to biochemical networks. [*de Lorenzo*, 2008; *Park et al.*, 2008; *Yuan et al.*, 2008; *Young et al.*, 2008; *Zhu et al.*, 2008; *Feist et al.*, 2009; *Zak and Aderem*, 2009].

With the increased usage of high-throughput technologies, the statistical analysis requires appropriate bioinformatical workflows. The intention of applying several bioinformatical approaches is to understand cancer as an integrated system of genes and protein, networks, and interactions.

This work focuses on investigating cellular networks, which represent signaling pathways implicated in many cancers types. For that, two projects were investigated within the scope of this thesis: the first study investigates signaling pathways in lymphomas and the second, signaling pathways in breast cancer. Each project consists of a coupled data set of gene and protein high-throughput measurements. Methods for network reconstruction are applied to each data set and combined with existing biological knowledge, e.g. signaling pathway (from databases KEGG, Reactome, NCI and Biocarta).
Both projects address the same two main questions:

(i) Are our methods suitable to reconstruct existing knowledge?

(ii) Can new edges in the networks be identified that explain the interaction of key pathway members?

The purpose of the lymphoma project was to widen the analyses of signaling in B cell lymphoma by looking at different data sources: transcriptome (section 2.5.1.1) and phospho-proteome data sets (section 2.5.1.2) using different network reconstruction techniques. The results of this analyses are reported in the section 3.1.

The main focus of the second study lied on the role of ROR2 in WNT11 signaling in breast cancer. Here again, different data sources (RNA-seq with RPPA) were used. This time transcriptome (section 2.5.2.2.1) and (phospho-)proteome (section 2.5.2.2.2) time series data sets were examined to study pathways at gene and protein level. The results of this analyses are presented in the section 3.2.

In the 4.1 *Discussion* chapter, challenges, which arose within the lymphoma cancer study (section 4.1.1) were discussed. The final discussion section 4.1.2 discusses the results of the WNT11 signaling network reconstruction in ROR2 overexpressing breast cancer cells.

CHAPTER II

Material and Methods

2.1 Methods for the statistical analysis of transcriptome and (phospho-)proteome data

Biological data are sensitive to different noises and errors, consequently a number of steps are necessary to pre-process raw measurements. Due to the presence of various technical variability, it requires normalization of intensity measurements of all platforms to remove systematic biases. The resulting pre-processed data consist in corrected and normalized raw data that can be further statistically analyzed to investigate expression levels in different sample groups. Approaches for pre-processing rely on the type and structure of data. Methods for microarray data are, for example, different from that for proteomic data.

As this is not the main focus of this thesis, a detailed description is provided. A general and comprehensive explanation for microarray data can be found in [*Yang et al.*, 2002; *Irizarry et al.*, 2003] and for RNA-seq data in [*Zyprych-Walczak et al.*, 2015] .

Analysis of Affymetrix microarray and Luminex xMAP data

The raw microarray dataset and Luminex xMAP phospho-proteome dataset were normalized using quantile normalization to make the distributions the same across samples. For this step, the *normalizeQuantiles* function implemented in the R/Bioconductor package *limma* was used. [*Smyth*, 2005] Afterwards, the normalized values were transformed into $\log_2$-scaled expressions. Within preprocessing, steps before the main (differential) analysis, probes, that could not be mapped onto any Entrez Gene ID, were removed. Then, differentially expression values were calculated using a linear fit model and an empirical Bayes method in the limma package.

Analysis of RNA-seq data

RNA-seq data were first quality checked via FastQC [Babraham Bioinformatics, *Andrews* [2010]] and then aligned to the transcriptome using STAR tool [*Dobin et al.*, 2013]. Gene-level abundances were estimated using the RSEM algorithm [*Li and Dewey*, 2011]. Further pre-processing steps were done using *edgeR* [*Robinson et al.*, 2010] R package. Within the pre-filtering, rows in which there are very few reads were removed and genes that have at least 10 reads for some samples were kept as described by *Chen et al.* [2016]. DEG tools provide a way to estimate the read count differences between the conditions for every gene. Differentially expressed genes between different conditions were analyzed by fitting linear regression models, which usually take the following typical form: $Y = b_0 + b_1 x + e$. Here, Y involves all read counts from all conditions for a given gene. b_0 is called intercept and x is the condition. In the context of RNA-seq, it is very often a discrete factor, for example, treatment or control. e is a term capturing the error or uncertainty, and b_1 is the coefficient that captures the difference. *edgeR* fits negative binomial generalized linear models to every single gene. [*Robinson and Smyth*, 2008]

Analysis of RPPA data

The first step in the analysis of the RPPA data was to oversee the quality. The quantile-quantile plots of the serial dilution [*Zhang et al.*, 2009] were employed as a visual instrument for each slide manually and sorted out measurements with a controversial dilution curve.
To correct the foreground expression data to the dilution intercepts, the *correctDilinterc()* function of the R package *RPPanalyzer* [*Mannsperger et al.*, 2010] was used. This function removes the local background intensity at one spot from its foreground intensity.

In the second step, the background corrected data was normalized. This is a crucial step in RPPA data analysis to ensure sample comparability. To perform a spot-specific normalization of the signal intensities the *normalizeRPPA()* function of the R package *RPPanalyzer* [*von der Heyde et al.*, 2014] was applied. Hereby each array is normalized through his corresponding array on the normalizer slide.

After the pre-processing differentially expressed proteins between different conditions were also analyzed by fitting negative binomial generalized linear models. [*Robinson and Smyth*, 2008]

The estimated p values from the analyses of all four platforms were adjusted for multiple testing using the method of [*Benjamini and Hochberg*, 1995] resulting in FDR.

2.2 Methods for Reconstructing and Visualizing Network

I'm going to provide the reader an overview of the investigated network models. At first, Bayesian models in general and two different kind of network approaches which belong to the class of Bayesian network methods are presented. Afterwards, a literature based integration method is introduced. In the end of this section the utilized visualization techniques to create the learned networks are explained.

2.2.0.1 Bayesian Networks

Bayesian networks belong to the group of probabilistic graphical models (GM). They are mathematically precise and instinctively understandable to combine network analysis with Bayesian statistics. The graphical structures are used to represent knowledge about an unclear field. For instance, each node in the graph represents a gene or protein, while the edges between these nodes symbolize probabilistic dependencies among the corresponding gene or protein. Principally, BNs are a special case of the GM structure named directed acyclic graph (DAG). The structure of a DAG is defined by two sets: the set of nodes (vertices) and the set of directed edges pointing in the direction of influence. The advantage is that they enable a direct representation of the joint probability distribution (JPD) over a set of variables. [*Pearl*, 1988] They can be used to learn causal relationships and gain an understanding of the various problem domains.

An edge from node X_i to node X_j symbolizes a statistical dependence between the corresponding variables, roughly speaking that variable X_i 'influences' X_j. For example, a BN could represent the probabilistic relationships between a set of genes. Given the activation or inhibition of a specific gene, the network can be used to estimate the probabilities of activation or inhibition of a different gene and so represent a signaling flow in a cell.
As pointed out by [*David*, 1999], this construction is optimal for incorporating prior knowledge whenever available.

Most of the times the BN is unknown and needs to be learned from the data. This question is referred as a learning problem, which can be described generally in this way: Given a data set and prior information (e.g., expert knowledge and literature) estimate network structure and the parameters of JPD in BN

$$P(X) = \prod_{i=1}^{n} = P(X_i = x_i | pa(X_i)) \tag{2.1}$$

with parent gene pa(X_i) as a regulator of the gene i, where the probability is conditioned.
Bayesian networks fill the local Markov property, which states that a node is conditionally independent of its non-descendants given its parent nodes and thus JDP can be written as a product of conditional probabilities.
A scoring metric is applied to assess the model. The objective here is to infer a

network model that represents the data with high probability. One of the most popular scores is the Bayesian Information Criterion (BIC) [*Schwarz*, 1978] which also penalizes graph complexity to avoid overfitting.

2.2.1 Nested Effects Model

In this section, the idea and initial definition of Nested Effects Model (NEM), as introduced in [*Markowetz et al.*, 2005; *Tresch and Markowetz*, 2008], are briefly explained.
The main idea of NEM is that perturbing genes at the beginning of a signaling pathway will affect all targets of the transcription factors while perturbing a single downstream TFs will only affect its direct targets. These direct targets represent just a subset from the phenotypes observed after disturbing the entire pathway.
This leads to a nested structure of affected gene sets located downstream in the pathway.

Following the NEM literature, NEM distinguishes between silenced genes (S-genes) and genes showing a downstream effect (E-genes). That means genes with a high-expression change are identified as E-genes. In each experiment, one S-gene is silenced and the effects on E-genes are measured by microarrays. Each S-gene needs to be silenced at least once and S-genes and E-genes can but do not have to overlap.

The original approach from [*Markowetz et al.*, 2005] performs at first a discretization step on a count matrix, which contains the counts of how often a specific gene shows an effect. The discrete values 0 and 1 indicate if a disruption of signal flow was detected or not.
Later, several extensions were published. Firstly, [*Fröhlich et al.*, 2007] overcome to discretize the data, they calculate the p value distribution of the differential gene expressions. They also introduced the inference approach called module network, which assembles the final network recursively from smaller subnetworks.
To allow the integration of prior assumptions another enhancement is provided by *Zeller et al.* [2009], which brings the original approach in the Bayesian environment.

Next, a brief overview of how the signaling schemes are inferred is presented.
Given a set of E-genes $E = \{E_1, \ldots, E_m\}$, and a set of S-genes $S = \{S_1, \ldots, S_n\}$ a pathway model, it is assumed as a directed graph T on vertex set S as a starting point.

The subset of S-genes is interpreted as 'influence region of S'. All influence regions together form the 'silencing scheme Φ', which is stored as an adjacency matrix $\Phi \in \{0, 1\}_{n \times n}$.
The concept is that intervention at a singular S-gene puts this its state to 1. The silencing effect is propagated along the directed edges of T.
Then the extended graph to $T' = S \cup E$ encodes the connection between each E-gene to its S-gene. In general, every E-gene has a single parent in S, but if more than one

S-gene regulates an E-gene, the average over the observed effects is taken. The state of E-genes can be 0 or 1 and whether their parent S-gene is in the influence region or not. This state is drawn from the microarray measurements.

After the discretization of input expression data, the algorithm scores through a list of pathway hypotheses. Each hypothesis predicts downstream effects at E-genes. According to the score, the algorithm sorts the silencing schemes how well potential pathways fit experimental data. The predicted expected effects can be compared with observed effects to choose the silencing scheme, which fits the data best.

In the following, the marginal likelihood introduced in *Markowetz et al.* [2005] is described. The likelihood of the data is a product of the probabilities of observing or missing an effect over all E-genes. Note that the 'true' T' of a candidate graph T is unknown for two reasons: (i) the positions of E-genes are unknown and (ii) they can be regulated by more than one S-gene. Also, only the graph T of S-genes is of interest and not in the position to E-genes. To overcome these points they calculate the marginal likelihood by average over the edges between S- and E-genes.

Fröhlich et al. [2007] extended the approach to a more general inference scheme. Instead of counts, they deal with a matrix of (raw) p values. These p values specify the likelihood of a gene *a* if it is differentially expressed after knock-down of gene *b*. This overcomes the critical discretization procedure, which has a direct influence on the conditional likelihood and can be difficult to estimate.
Additionally, the p values are fitted using a so-called three component beta-uniform mixture model (BUM) consisting of a uniform and two beta distributions. [*Fröhlich et al.*, 2007] could show an improvement in fitting the p values using this modification.

Furthermore, they apply a Bayesian prior reflecting the degree of belief in the existence of edges in the network. The smaller this difference to the prior assumptions, the higher the edge probabilities should be. Therefore Laplacian distribution is an appropriate model to characterize the probabilities.

Markowetz et al. [2005] completely score all possible topologies, which is just applicable for very small networks. To overcome this issue *Fröhlich et al.* [2007] introduced a heuristic called module networks that evolves a graph from subgraphs, named modules.

First of all, they calculate the hierarchical clustering of the expression profiles. This realizes the assumption that S-genes with a similar E-gene response profile are close neighbors in the pathway. After estimating all clusters (modules), the algorithm composes their connections. To find this connection the method uses the greedy hill-climbing algorithm. Edges between S-genes will be added subsequently to the complete network if they increase the likelihood.
This algorithm computed much faster and therefore allowed for the inference of large-scale networks compared to the original approach. [*Fröhlich et al.*, 2007]

NEM needs high-dimensional, indirect measurements of rather qualitative knockdown effects, such as microarrays and is not able to model the time-dependent behavior of the system.

2.2.2 Dynamic Deterministic Effect Propagation Network

In this section, an overview of the general DDEPN framework, which was applied to the phospho-protein measurements, is given. DDEPN is a network inference method for high-throughput data with direct observation of involved proteins after knockdown of the measured components.

As a first step, DDEPN identifies the state transitions of a hidden Markov model, where the states correspond to combinations of activities of nodes in the network.
In the second step, a likelihood function scores the candidate networks derived from the previously estimated state transitions.

The nodes represent the measured proteins. The signal flow through a given network of proteins is represented in a matrix, which contains a series of possible system states. DDEPN treats each perturbation as an external influence and includes it as a node into the network with the constantly active state.
The type of each edge is stored in an adjacency matrix giving 0 for no edge, 1 for activation and 2 for inhibition. The algorithm starts at the stimuli nodes and then resolves the status of all children. For example, a child becomes active if all parents, who are connected via inhibition edges are inactive and at least one parent, who is connected by an activation edge has to be active.

The method continues with the Viterbi training algorithm to find a series of reachable system states that are agreeing with the measured experimental data. This algorithm starts with sampling random states and estimates model parameters depending on them. With these parameters new the system states are estimated using Hiden Marcov Model (HMM). This procedure iterates until convergence. After reaching convergence the likelihood is calculated for the resulting state matrix.

Bender et al. [2010] proposed that each measurement comes from different normal distributions whether the state of the individual protein is active or inactive. Thus there is an 'active' normal distribution if its state is 1, and from a 'passive' normal distribution, if its state inactive. The parameters for the distributions are calculated as unbiased empirical mean and standard deviation of all measurements for this protein in the given class.
In the next step, the algorithm uses GA to search through the whole population of possible networks the optimal network structure. Hereby the BIC is used as a fitness score because it penalizes a higher number of edges.
At the end of this step, the final network is drawn from a combination of all candidate networks. Each edge that occurs in more than a defined fraction in the population is

present in the final network.

Within this approach, it is also possible to include prior knowledge. In DDEPN it is implemented via Bayes theorem:

$$P(\Phi) = \frac{P(D|\Phi)P(\Phi)}{P(D)}, \tag{2.2}$$

where Φ is the matrix of model parameters and D consists of measurements. $P(D|\Phi)$ is the likelihood function as defined in [*Bender et al.*, 2011], $P(\Phi)$ defines the prior distribution, and $P(D)$ is a constant normalizing factor.

2.2.3 Literature-based Data Integration

The cross-platform analysis intents to link biological conclusions on different signaling levels. There are integration methods which concern naive weighted means of transcript and protein abundances [*Balbin et al.*, 2013] or prize-collecting Steiner tree formalism (PCST) to find an optimum tree [*Tuncbag et al.*, 2013] and methods which consider consensus pathways and molecules [*Wachter and Beißbarth*, 2015]. Alternative con-cepts exploit the relationships between gene products to produce a network of related genes, known as an interactome. [*Gibbs et al.*, 2014] To elucidate a subnetwork or pathway containing gene products with functional relatedness *Dutkowski et al.* [2013] developed a de novo clustering algorithm of interactomes. On the other hand, some techniques integrate data sources before clustering applying a joint latent model. [*Shen et al.*, 2009; *Michaut et al.*, 2016]

The approach that is used in this work is called pwOmics. [*Wachter and Beißbarth*, 2015] This integration method takes the different molecular layers into consideration. There-fore public signaling pathway information and transcription knowledge data is used to identify molecular interactions.

Different pathway databases are used to classify the pathways of the differentially expressed (phospho-)proteins together with genes or transcripts in the down- and in the upstream analysis.

In the first place, the method analyzes the two data sets separately. This permits a level-specific interpretation of down- and upstream changes of regulatory molecules in each inhibition experiment. The following analysis steps deal with pre-processed transcriptome and (phospho-)proteome data. That means the integration approach takes already normalized and perhaps filtered data sets as input.

To identify downstream target genes, upstream TFs and later proteomic regulators it is possible to use different pathway databases. A number of public databases systematically gather pathway information. The focus of this package lies on four databases that supply their data in the Biological Pathways Exchange (BioPAX) format: KEGG [*Kanehisa and Goto*, 2000], Reactome [*Croft et al.*, 2014], Pathway

Interaction Database [*Schaefer et al.*, 2009] and Biocarta [*Nishimura*, 2001].

The downstream analysis is motivated by assuming that protein phosphorylation transmits downstream regulation. In other words, pathways, which include differentially abundant (phospho-)proteins, are determined. Then, gene sets of the identified pathways are matched against TFs derived from the TF-target gene database. In the same way, downstream target genes are identified. A target gene is a gene regulated by a given TF.

To replenish this analysis step the upstream analysis of the transcriptome data set provides TFs and proteomic regulators based on differentially expression levels. The purpose here is to detect pathways containing transcripts of possible upstream proteomic regulators. Thus, firstly upstream TFs of significantly differentially regulated transcripts are identified and afterward, pathways they belong to are determined. In order to ensure matching genes and proteins, their IDs are transformed into HUGO gene symbols.

Following the pwOmics literature, the individual analysis follows the static consensus analysis where signaling networks were constructed depending on intersecting proteins, TFs, genes and transcripts on each cellular layer. Consequently, the results derived from individual downstream and upstream analyses are reduced to molecules that are identified from both platforms. The corresponding proteins and TFs are mapped to the PPI STRING database [*Franceschini et al.*, 2013]. Steiner trees are generated using a variant of the shortest paths algorithm [*Sadeghi and Fröhlich*, 2013]. Steiner nodes are inserted to assure the connectivity of the network.
Finally, matching TF-target interactions are added by integrating TRANSFAC® [*Matys et al.*, 2006] information. The accomplished networks arrange interaction and regulatory information on the consensus molecules.

2.2.4 Network Visualization with R

Visualizations are essential to retrieve fast and easy access to different aspects of the network because the interaction between biological components can not only be measured experimentally but also calculated. Several approaches to visualize biological pathways and relations are possible and offer particular concentration to features. A more detailed mathematical discussion of graphs within network biology can be found in [*Emmert-Streib and Dehmer*, 2011].

Though not specifically developed for it, R has become a powerful tool for network analysis. To investigate network properties, the R package used in this thesis is called *igraph* [*Csardi and Nepusz*, 2006]. It has implemented considerable tools to visualize network structures. *igraph* handles both undirected and directed graphs. Within the package, different algorithms are provided that allow graphs to be displayed in an assortment of layouts. The *igraph* package can deal with labeled and unlabeled and weighted and unweighted networks. It also supports simple graph-theoretical methods and some basic network descriptors to define basic structural features (e.g., degree and coefficient of global clustering).

In this thesis, each node of a network is a gene or, respectively, a protein and each edge indicates whether there exists a dependency between them, whereby the direction of an edge indicates the orientation of influence.

2.3 Survival Analysis

As part of breast cancer patient data analysis, a Kaplan-Meier (KM) analysis was performed. The KM method is a non-parametric estimator to estimate and graph survival probabilities as a function of time. To determine any statistical difference between the survival curves the log-rank test, implemented in *survival* R-package [*Therneau*, 2015], was used.

In principle, the log-rank test uses the total number of deaths reported and the total expected number of deaths in each group to generate a test statistic. This test statistic is later analyzed using the χ^2 test with the null hypothesis assuming that all survival curves are the same. Survival analysis can be applied to any event of interest. In this thesis, it was used to estimate the survival function of overall survival (OS) and metastasis-free survival (MFS) among the primary and metastatic tumor of breast cancer. The KM curves were compared using a log-rank test implemented in *survival* R-package. As part of the analysis of breast cancer patient data, full hierarchical clustering was performed using Pearson correlation as a distance measure.
To divide the different patient groups into clusters within the dendrogram, the *cutreeDynamic* algorithm, as implemented in the *dynamicTreeCut* R package [*Langfelder et al.*, 2007], was used. Subsequently, the identified patient clusters were investigated regarding their MFS.

2.4 R packages

All analyses steps were done in R version 3.4.4. The aforementioned methods are implemented in different R packages. An overview of all R packages that are used within this thesis is given in table 2.1. The table lists the packages, a short description and the repository, from where they can be installed.

Table 2.1: *Table of the utilized R packages.*

Name	Version	Description	Repository
RPPAnalyzer	1.4.5	Reading and Normalizing RPPA Data	CRAN
edgeR	3.20.9	Differential expression analysis of RNA-seq expression profiles	Bioconductor
limma	3.34.9	Differential expression analysis of micro-array data	Bioconductor
NEM	2.52.0	Network Reconstruction of readouts from perturbation experiments	Bioconductor
DDEPN	2.2.3	Infering signaling networks for time course data	CRAN
pwOmics	1.10.1	Pathway-based data integration of omics data	Bioconductor
upsetR	1.4.0	Visualizing the intersection of sets	CRAN
ggplot2	3.2.1	Data visualization	CRAN
igraph	1.2.4.1	Functions for network analysis and visualization	CRAN
survival	2.44.1	Functions for survival analysis	CRAN
dynamicTreeCut	1.63.1	Identifying clusters in hierarchical clustering dendrograms	CRAN

2.5 Data Sets

2.5.1 Lymphoma

To describe the dominant pattern of gene and protein expression and to derive pathway activity in BL from measured (global) expression changes, the BL2 cell line was used as the model system. One way to overcome the limitations of learning networks from single data types is to use publicly available sources of complementary data. The data sets investigated in this project includes public available gene expression measurements and phospho-protein levels generated in the network analysis of BCR signaling. Further details, as well as the pre-processing steps, performed on both microarray raw data and proteomic raw data, are described in the following section.

2.5.1.1 Gene expression data

Gene expression measurements by microarrays are very widespread as a data source to allocate gene functions because they provide a comprehensive picture of gene activity in cells.

Within this thesis work, a public microarray data set, that provides measurements of gene expression in the Burkitt lymphoma cell line BL2, was used. The raw microarray data files are available at the NCBI Gene Expression Omnibus (GEO) [*Pirkl et al.*, 2016] database under the accession number GSE68761. The BCR data was generated using Affymetrix GeneChip Human Genome U133 Plus 2.0 arrays. It provides gene expression measurements after the perturbation on 5Z-7-Oxozeaenol (TAK1), IKK2 inhibitor VIII (IKK2), Ly294002 (PI3K), SB203580 (P38/MAPK14), SP600125 (JNK), U0126 (ERK1/2).

2.5.1.2 Phospho-protein data

The phospho-protein expressions, again, in the cell line BL2 were measured using Luminex bead-based multiplex assays. The cells were stimulated with the growth factors IgM for 30 min to activate the BCR receptor. Afterward, the cells were preincubated for 90 min with pharmacological inhibitors against the kinases BTK (Ibrutinib), MEK (AZD6400, U0126), PI3K (BMK120, Ly294002, Cal-101), AKT (MK-2206), TAK1 (5Z-7-Oxozeaenol), IKK (ACHP, MLN120B), JNK (SP600125, JNK inhibitor VIII), p38 (SB203580) or mTOR (Rapamycin).

While being aware of having a biased selection of phospho-proteins, the assay included proteins that are within or nearby the stimulated pathway and the inhibited kinases. Read outs of 15 phospho-protein levels were done 90min after inhibition to reach a network to be approximately in a steady state. That means that, after treatment and incubation, lysates were assembled and analyzed with the BioPlex Protein Array system (BioRad, Hercules, CA) using beads specific for pAKT (S473), pERK1/2 (Thr202/Tyr204/Thr185/Tyr187), pGSK3α/β (S21/S9), pMEK1 (S217/S221), pJNK (T183/T185), pSYK (Y352), pBTK (Y223), pZAP70 (Y319), pNF-κB (S536), pP38 MAPK (T180/Y182), pJUNC (S63), pHSP27 (S78), pRPS6 (S235/S236) and pBAD (S136).

The data set for this project was generated by Anja Sieber of the Institute of Pathology (Charité University Medicine Berlin) as published previously [*Klinger et al.*, 2013]. The BioPlex manager software was used for data acquisition. Figure 2.1 gives an overview of the whole experimental setting of the project.

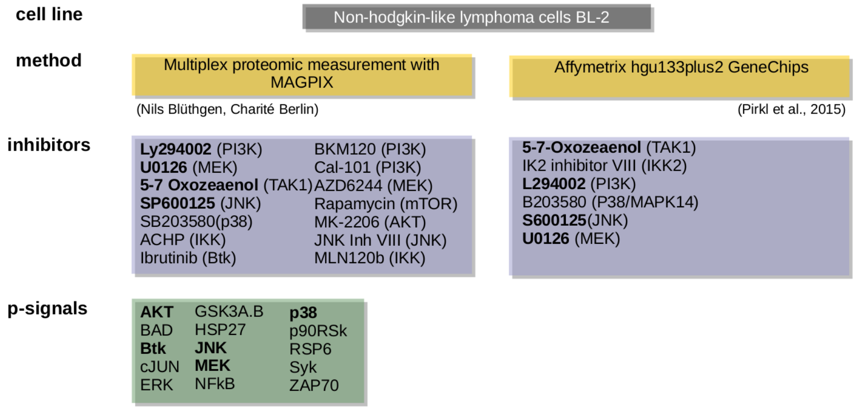


Figure 2.1: *Overview about the whole experimental setting. Bold proteins represent the intersection of inhibitions in both data sets and measured proteins that where also inhibited in one experimental condition.*

2.5.2 Breast Cancer

Although many of the members of WNT11 signaling pathways are now known, much remains to be discovered regarding specific receptor-ligand pairings. For instance, receptor tyrosine kinase like orphan receptor 2 (ROR2) is a WNT11 receptor that has been shown to be highly expressed in breast cancer metastases. [*Klemm et al.*, 2011] The study from *Bayerlová et al.* [2017] indicated that ROR2 mediated WNT11 signaling plays a central role in breast cancer progression. The following work aims to further investigate the role of ROR2 in WNT11 signaling.

2.5.2.1 Experimental set up

To investigate the role of ROR2 in WNT11 signaling, in vitro experiments for data generation were performed in cooperation with the research group of Prof. Annalen Bleckmann (University Hospital Münster, Department of Internal Medicine-A). As model cell line hormone receptor-positive MCF-7 human breast cancer cells (DSMZ, Braunschweig) were used. Cells either expressed pRor2 or the respective pcDNA3.2 empty vector (as in *Bayerlová et al.* [2017]) and were kept under constant antibiotic selection with G418 (750g/ml, Roche) or zeocin (10g/ml, Invitrogen).
For the knockdown of WNT11 gene expression, MCF-7 cells were transfected in suspension with RNAimax reagent (ThermoFisher Scientific) using 10nM control siRNA (siCTL, #sc-37007) or a siRNA pool directed against Wnt11 (siWnt11, #sc-41120, both santa cruz). Cells were used 48h post transfection for subsequent experiments.
For stimulation experiments, MCF-7 cells were treated with rhWnt11 (100ng/ml, #6179-WN/CF, R&D) for the indicated time periods. The laboratory implementation was realized by Lena Ries (University Medical Center Göttingen, Department of Hematology/Medical Oncology)

The following Figure 2.2 shows the experimental setting, which was designed by Astrid Wachter (University Medical Center Göttingen, Department of Medical Bioinformatics) and Kerstin Menck (University Hospital Münster, Department of Internal Medicine-A).

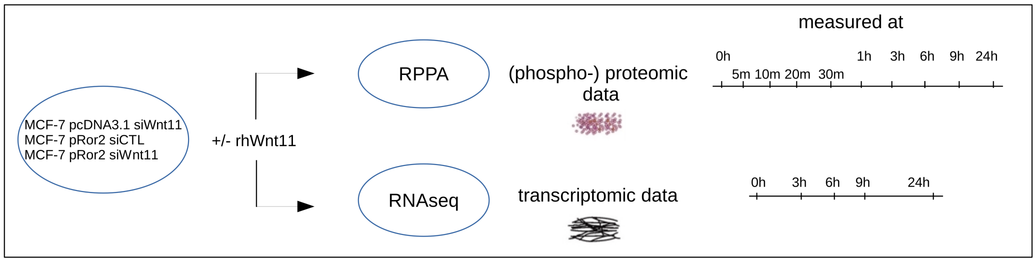


Figure 2.2: *Overview about the whole experimental setting.*

2.5.2.2 Gene and protein expression data

2.5.2.2.1 Gene expression data

Newly generated RNA-seq time series data set

To quantify the gene expression changes linked with the observed interplay of WNT11 and ROR2 in MCF-7 cells, three replicates for each experiment were executed using RNA-seq. Cells were treated as described in 2.5.2.1 and then were sequenced at NGS Integrative Genomics Core Unit (NIG) by Dr. Gabriela Salinas-Riester. The measuring points cover 0, 3, 6, 9 and 24 hours after stimulation. The reads were then mapped against the reference genome GRCh38 using database information from Ensembl ver. 38.78.

Publicly available RNA-seq data set

To associate WNT11 signaling with clinical outcome, RNA-seq data from brain metastatic tissue published by *Blazquez et al.* [2018]. This data set contains total RNA reads of 48 breast cancer patient tissue and overall survival annotations, from the date of brain metastasis resection forward.

Microarray data set

To examine the role of selected WNT11 pathway members in the development of metastasis in breast cancer patients, a data set of 2075 patients with breast cancer was included in the study. The data set is a compilation of ten gene expression data sets of primary breast cancer patient samples. This data set was compiled by *Bayerlová et al.* [2017].

2.5.2.2.2 Protein expression data

Newly generated RPPA time series data set

The newly generated RPPA data set for this project was generated by Eileen Reinz of the Division of Molecular Genome Analysis in Heidelberg headed by Prof. Dr. Stefan Wiemann in the group of Dr. Urlike Korf. MCF-7 Cells were treated as decribed in 2.5.2.1. The RPPA chip covered measurements of 67 proteins and 34 phosphoproteins. Expression levels were analyzed for each treatment over ten different time points. For the short-term measurements were carried out at six time points (0min, 5min, 10min, 20min, 30min, 60 min) were measured, for the long-term measuring four time points (3h, 6h, 9h, 24h). The experiment was carried out with three biological replicates.

Publicly available RPPA data set

To study the protein expression in breast cancer patients, a publicly available data set was supplied to a survival analysis. The utilized RPPA data contains measurements from brain metastatic tissue of 48 breast cancer patients and was published by *Blazquez et al.* [2018].

CHAPTER III

Results

3.1 Lymphomas

This chapter is divided into four parts. First, an overview of the statistical analysis of the two data sets is presented. The second section summarizes the network reconstruction results of transcriptome data (section 2.5.1.1), followed by the study of the phospho-protein data (section 2.5.1.2). Finally, both data sets were combined and a literature-based integration analysis was performed.

3.1.1 Differential Expression Analysis

After normalizing the microarray data set as described in section 2.1, each inhibition experi-ment was compared against the control of the gene expression data. In Figure 3.2 B the resulting $\log_2$ transformed FCs of the 250 most significant genes (FDR < 0.05) are displayed. The significant DEGs were then considered in the subsequent network analysis.

Interestingly, inhibition of ERK and PI3K show comparable effects on the gene expression level. The PI3K and ERK pathways both are known as important intracellular signaling pathways. Oncogenic alterations of the effectors in PI3K and ERK pathways are frequently observed in many cancers. [*Kohno and Pouyssegur*, 2006; *Jokinen and Koivunen*, 2015] A broad crosstalk between these two pathways has been invested. The both pathways functionally co-regulate the same transcription factors which drive cell proliferation and cell survival. [*McCubrey et al.*, 2007; *Mendoza et al.*, 2011]

In the same way, the phospho-protein expression data described in section 2.5.1.2 was investigated. Before calculating $\log_2$ FCs of the phospho-proteins, the distributions of their expression values were plotted to see if a normalization step is necessary. Figure 3.1 shows the density plot and box plot for measured phospho-protein expressions, revealing that the distributions of all proteins are not skewed and have no or just a few outliers. Although some protein expression are closer to a bell shaped (normal) curve than others, in particular pGSK3α/β. For this reason, it seemed reasonable to include a normalization step before performing the statistical analysis.

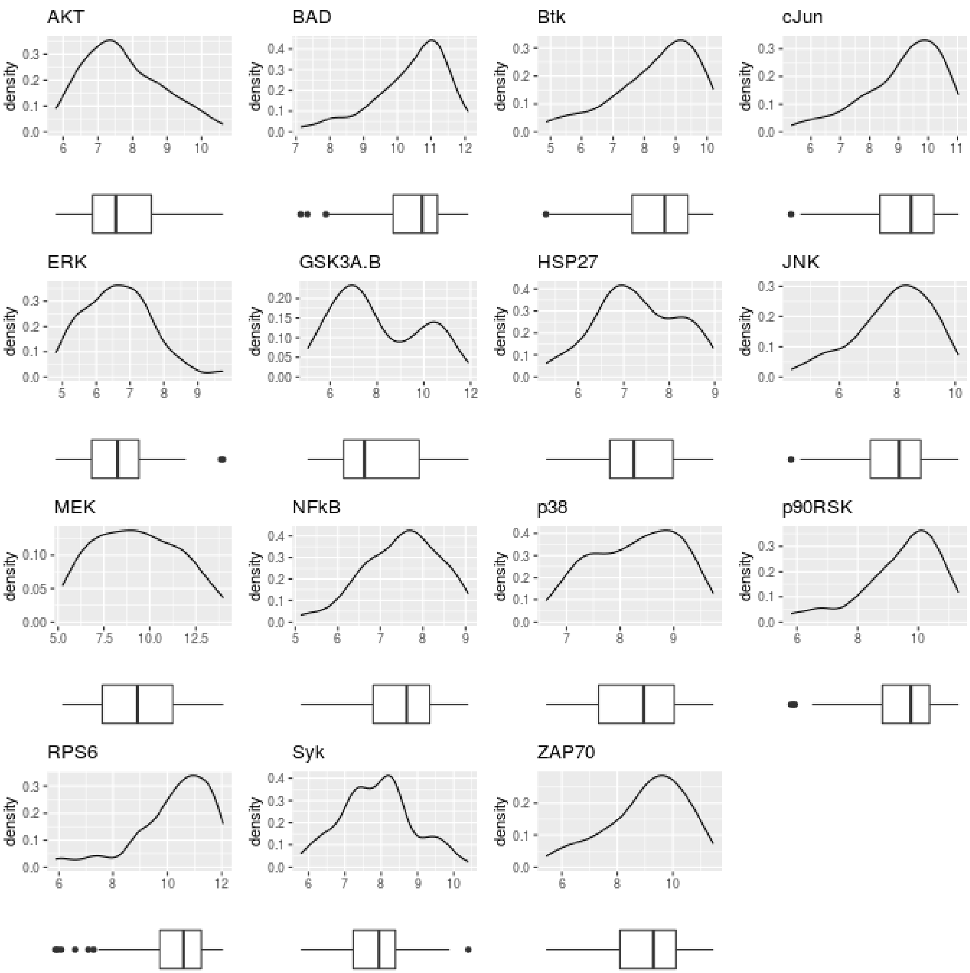


Figure 3.1: *Boxplot and density curves showing the $\log_2$ transformed protein expression levels for each phospho-protein separately.*

In Figure 3.2 (A), the result of DE analysis is shown as $\log_2$ transformed fold changes in the inhibition experiment compared with the unperturbed controls.

Nearly every phospho-protein shows very small $\log_2$ fold changes and seems to be slightly down-regulated after most interventions. As an exception, MEK reveals a strong effect after perturbation of TAK1, p38 MAPK as well as combined treatment of p38 MAPK and IKK. Furthermore, MEK inhibition leads to lower phosphorylation of SYK, ZAP70, RPS6, ERK1/2 and of the AKT, p38 MAPK, and JNK pathway, whereas its own phosphorylation was increased. Also, it is well known that ERK reduces its own activity through inactivating phosphorylations of RAF-1 and MEK [*Steelman et al.*, 2011]. This suggests that the signal inhibition of ERK1/2 increased the phosphorylation of MEK. Furthermore, the perturbation of p38 MAPK highly elevated the phosphorylation of MEK and ERK1/2, but only marginally increased phosphoryla-tions of the other proteins, namely ZAP70, BTK, AKT, GSK3A/B, JNK, c-JUN, and NF-κB.

Due to the low expression pattern in the phospho-proteins, none of them was significantly differentially expressed (FDR < 0.05). Nevertheless, it should be noticed that a statistically significant difference in the expression level does not imply the incidence of any difference in biological significance. That is because the mathematical definition of 'differential expression' as any non-zero difference does not correspond exactly to the differential expression biologists follow. Overall, these data reflect the general image of BCR pathway activation in B cell lymphomas but also indicate that more subtle systemic alteration underlie this dysregulation, instead of significant changes in the expression of one or more components. Consequently, all phospho-proteins were included in further analysis.

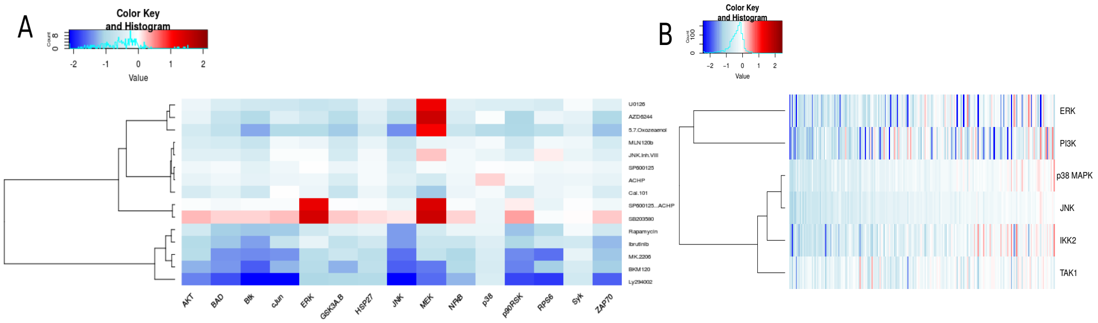


Figure 3.2: *Hierarchical clustering analysis and mean log_2 FCs of the A: measured phospho-proteins in the columns and inhibition experiments in the rows and B: 250 most significant genes in the columns and the inhibition experiments in the rows*

3.1.2 NEM

Since the utilized data sets are different in their measurement technique (e.g. Affymetrix mircoarray and Luminex protein assays), on each data set different network inference approaches were applied to estimate the signaling between genes and proteins respectively. To study and re-evaluate the microarrays data described in section 2.5.1.1, NEM approach (section 2.2.1) was applied. After filtering of FDR < 0.05, a three component beta-uniform mixture (BUM) model of the p values was fitted via an EM algorithm.

Subsequently, NEM was executed using a prior knowledge graph. The summarized graph for this analysis step, shown in Figure 3.4, focuses on the genes, which are S-genes in the microarray. Figure 3.3 shows a summary of a part of BCR signaling adapted from *Pirkl et al.* [2016].

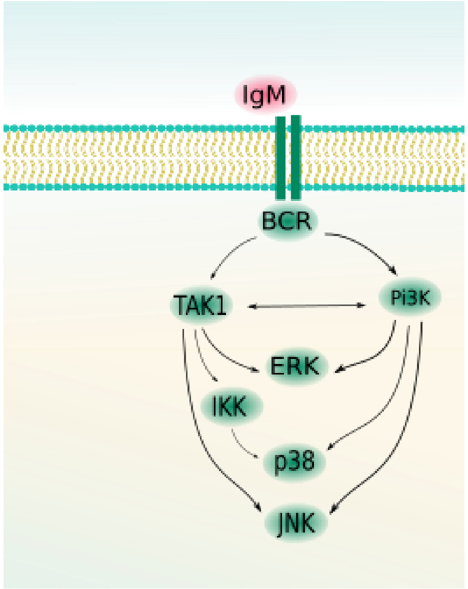


Figure 3.3: *Prior knowledge network for the NEM analysis adapted from Pirkl et al. [2016]*

The resulting model, shown in Figure 3.4, contains many new edges and only three edges, which overlap with the prior knowledge graph. Based on this gene expression measurements, the methods generates new hypotheses of signaling transduction. The learned network predicts that the activation of the JNK pathway is PI3K and ERK-dependent, while TAK1 relies only on ERK. The signal flow to p38 MAPK can be stopped with the inhibition of PI3K. The three kinases p38 MAPK, JNK, and IKK2 regulate each other and build a loop in this network. ERK propagates signals into the p38 MAPK and TAK1 pathway.

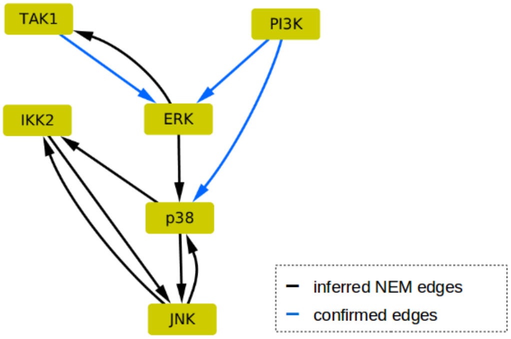


Figure 3.4: *The highest scoring network edges: The nodes in this graph represent genes, which were perturbed in the biological experiments, and the edges can be interpreted as the signal between genes. Blue edges represent confirmed interactions and black edges are newly inferred by NEM.*

The result presents a first model of the BCR signaling network and a hypothesis explaining how downstream nodes of this pathway could be affected. Nonetheless, this analysis could not perfectly explain the signaling network. To extend the network structure, phospho-protein network analyses were performed to reveal the downstream signaling flow.

3.1.3 DDEPN

In this step, the aim was the identification of individual drug response patterns in lyphoma cells. For Bayesian network reconstruction on the previously described phospho-protein data set (section 2.5.1.2) DDEPN approach was run. The resulting graphs are shown in Figure 3.6.
As mentioned in the first chapter, in the field of network analysis, it could be shown that integrating prior knowledge can enhance the result and it helps to decrease the search space. For these reasons, a prior knowledge graph composed of literature knowledge was used in this protein network analysis. Figure 3.5 shows a summary of recent knowledge of a component of BCR signaling, which was constructed manually from the pathway database KEGG. A detailed description of BCR signaling is given in section 1.1.1.

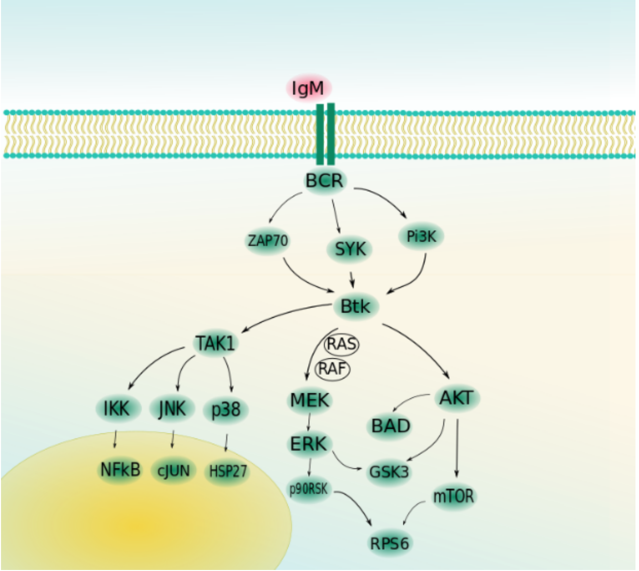


Figure 3.5: *Manually assembled graph representing the prior knowledge. The prior knowledge network was used as the starting point for the DDEPN model.*

The reconstruction started from predefined initial states of the network nodes. The inhibited nodes were set to specific values, reflecting the conducted experiment. The result of integrating the prior knowledge as explained in the previous section is given in Figure 3.6. Directed edges represent regulatory processes such as activation and inhibition. It can be seen, that DDEPN did not identify known interactions but infer some new edges between the proteins. Compared to the prior network, most of the newly inferred edges were direct connections between protein nodes. For instance, hte data does not support the chain from SYK over BTK, AKT, and mTOR to RPS6.

Instead the expression of the proteins was interpreted as a direct interaction between SYK and RPS6. The low, but rapid, phosphorylation of the proteins complicated the exact reconstruct the signaling pathway at it is known from literature.

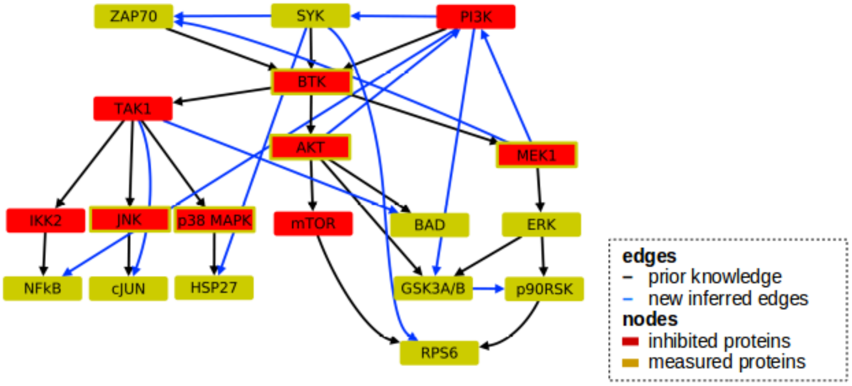


Figure 3.6: *The inferred BCR phospho-protein network using DDEPN approach. Comparison of the reconstructed network with prior knowledge.*

To compare the resulting NEM and DDPEN graphs and to get an overview of all the inferred interactions, both networks were merged into one summary network. Figure 3.7 shows that the graphs don't overlap, but complement each other. It is noticeable that the graphs are connected via the two nodes TAK1 and PI3K. Accordingly, there are no edges, that are inferred from both data sets. Although it would be interesting if interactions could be identified that are supported by transcriptome and proteome measurements.

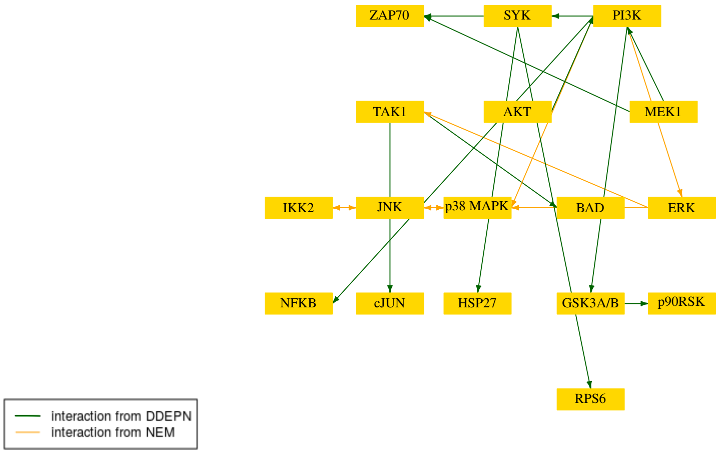


Figure 3.7: *Merging the resulting NEM and DDEPN graphs. Orange edges are derived from NEM analysis, and green edges are inferred by DDEPN analysis.*

Since the two reconstruction approaches are based on a different design, the resulting graphs share nodes. This restricts the reconstruction of one single network with edges that are inferred from both data sets in parallel. For the purpose of overcoming the missing overlap and to better integrate both data sets, a different network analysis approach was conducted in which knowledge from public literature sources was explored in a more comprehensive way. Specifically, the method *pwOmics*, which is explained in section 2.2.3, was utilized.

3.1.4 pwOmics

To perform the main integration steps, the open-source R package *pwOmics* [*Wachter and Beißbarth*, 2015] was used. Each step was performed for every perturbation experiment separately.

The first step encompassed a downstream analysis of the proteome data set. Pathway information in this procedure was taken from four databases KEGG [*Kanehisa and Goto*, 2000], Reactome [*Croft et al.*, 2014], Pathway Interaction Database [*Schaefer et al.*, 2009] and Biocarta [*Nishimura*, 2001]. Given that none of the phospho-proteins was significantly differentially expressed, the complete set was included in the input data for the downstream analysis. That resulted in a high number of pathways that are affected in downstream signaling.

The investigation of the four selected pathway databases in downstream analysis identified 270 pathways. Conferring to the number of TFs and targets in the downstream analysis the identified TFs activate the expression of a high number of genes.

For example, the EGFR signaling pathway was selected from this knowledge-driven model based on gene and phospho-protein expression data, respectively, for this condition. The EGFR signaling pathway is described to influence cancer progression in several cancer types. [*Wang et al.*, 2007; *Teixeira et al.*, 2009; *Costa et al.*, 2011] This indicates that the EGFR pathway is a relevant factor in cancer development and progression.

Table 3.1: *Result table of the downstream analysis.*

	IKK2	PI3K	JNK	p38	TAK1
No. of differentially abundant phospho-proteins	19	19	19	19	19
No. of pathways	270	270	270	270	270
No. of TFs	71	71	71	71	71
No. of potential target genes	874	874	874	874	874

For the upstream analysis, all transcripts were filtered with a p value below 0.05. Here, the TF-target gene interaction information is derived from the TRANSFAC database [Biobase version 2014.4; *Matys et al.* [2006]]. Like in the downstream analysis, the high number of differentially expressed genes resulted in a reasonable list of pathways. Identified upstream pathways included STAT3 dependent signaling pathway, which is involved in tumorigenesis. STAT3 is a transcription factor modulating many important functions in cellular transformation. [*Sansone and Bromberg*, 2012] The results for each of the five conditions are summarized in Table 3.2.

Table 3.2: *Result table of the upstream analysis.*

	IKK2	PI3K	JNK	p38	TAK1
No. of differentially expressed transcripts	1631	2299	902	963	641
No. of pathways	270	219	189	176	187
No. of TFs	412	466	339	307	621
No. of potential upstream proteomic regulators	980	816	676	1083	637

As reported by the TF-target gene database, the identified TFs activate the expression of a high number of proteins, as shown in Table 3.2.

In the static consensus analysis, the results of the different levels for each experiment were combined. Exemplary, the network of inhibition experiment of PI3K is shown in Figure 3.8 A.

The profiles of the consensus results display the existence of specific molecules in the consensus networks under each perturbation, as presented in Figure 3.8 B. Compared to the other conditions, after TAK1 inhibition, just a few consensus TFs could be identified by the algorithm. Subsequently, log_2 FCs were merged with the static consensus profiles reflecting the up or down-regulation of specific molecules in the consensus networks for each inhibition, as illustrated in Figure 3.8 C. As this figure implies, there are no down-regulated TFs. However, it was recognized that all TFs show small log_2 FC values. Further, it is noticeable that most effects are detected after PI3K inhibition. This is in line with recent studies who indicate a crucial role of the PI3K pathways in DLBCL [*Erdmann et al.*, 2017; *Bojarczuk et al.*, 2019] as well as an impact on B cell, which is strongly dependent on BCR expression [*Werner et al.*, 2010].

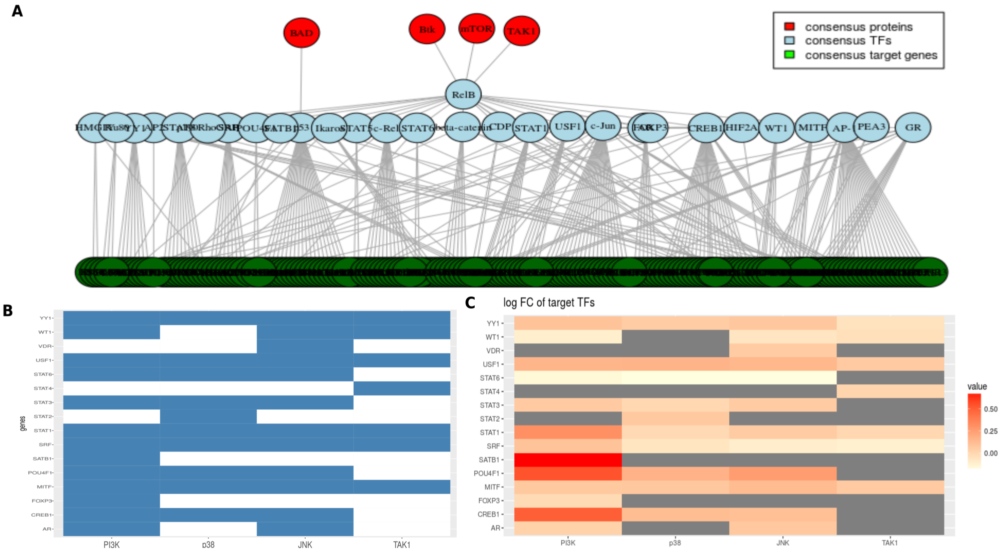


Figure 3.8: *Static consensus analysis results. A: Static consensus graph of PI3K inhibition. Red nodes are consensus proteins, blue nodes are consensus TFs and green nodes are consensus target genes. B: Consensus profiles for TFs of the different inhibitions. Rows represent the consensus TFs and columns show the conditions. Blue indicates that the gene is consensus gene under this condition, and white boxes indicate, that the gene is not in the set. C: Consensus map showing the $\log_2$ FCs of the TF. Rows represent the consensus TFs and columns show the conditions. Grey boxes indicate that the TF doesn't belong to the consensus graph under this condition.*

Approximately one third of consensus TFs appeared under all perturbation experiments. Among this consensus TFs, Serum Response Factor (SRF) was identified in every condition. It is a downstream target of a few pathways, such as the mitogen-activated protein kinase pathway (MAPK) and MEK/ERK pathway. SRF plays an important role in the regulation of proliferation and cytokine production in lymphocytes. [*Hao et al.*, 2003]
A second target, which dominated every consensus graph, even though it was not on the microarray, as c-Rel. c-Rel is a member of NF-κB transcription factor family. High c-Rel expression levels are described mostly in B and T cells, where many c-Rel target genes are associated with B and T cell malignancy. [*Gilmorec and Gerondakis*, 2011] In addition, one study indicates that c-Rel may also be directly involved in regulating DNA replication. [*Ishikawa et al.*, 1993]
Several genes were exclusively found only after one perturbation. A notable example is the vitamin D receptor gene (VDR). VDR polymorphisms are associated with an increased risk of BL. The findings of *Gascoyne et al.* [2017] indicate that the inhibition of VDR pathway activity may be of therapeutic benefit. Overall, these cases support the view that the inhibited targets have an important effect on downstream genes.

The pattern of the consensus target genes over all four conditions is displayed in

Figure 3.9. The heatmap shows genes with the highest absolute log_2 FC under the condition if inhibited PI3K. Grey boxes indicate, that this gene is not a consensus gene after that perturbation. It is noticeable, that most genes, which are in the set of consensus genes under PI3K perturbation, are not in the other consensus sets. The shown genes represent the over all pattern, that the four conditions don't share many consensus genes. A comparison of the four gene sets can be found in Figure 3.10.

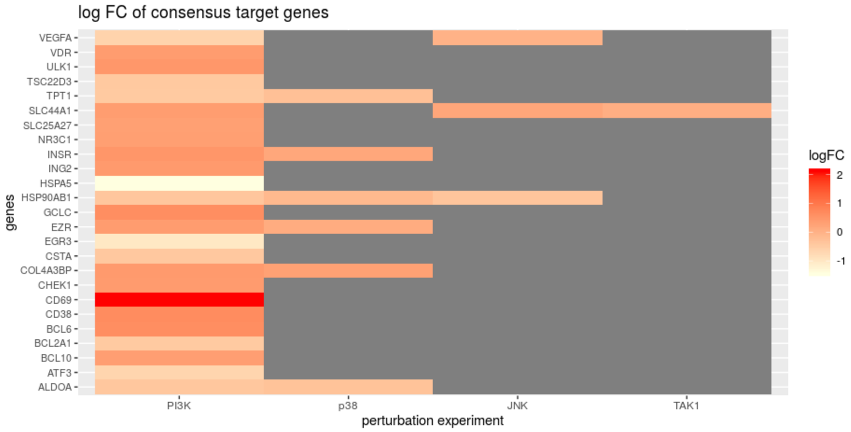


Figure 3.9: *Heatmap of log_2 FC of consensus target genes. Grey boxed indicated that the gene wasn't identified by the algorithm under that specific condition as a consensus target gene..*

Under PI3K inhibition, the CD69 gene showed the highest log_2 FC compared to all other log_2 FCs. CD69 is a marker expressed on the surface of activated leukocytes by activation of RAS, RAF and calcium release. [*D'Ambrosio et al.*, 1994; *Taylor-Fishwick and Siegel*, 1995] *Erlanson et al.* [1998] investigated its expression in B cell non-Hodgkin's lymphomas. They could show that 53% of aggressive cells expressed the CD69 antigen and demonstrated its impact on advanced stage and shorter survival.

To further study the pattern of targeted genes over all four conditions, the overlap between them was analyzed. The sets and their intersections are shown as a bar chart. The first four sets are mutually exclusive. That means that the bars include all the targeted genes that occur only in an individual set. The diagram shows that the sets of targeted genes are rather distinct. The vast majority of target genes is not shared between all four conditions.

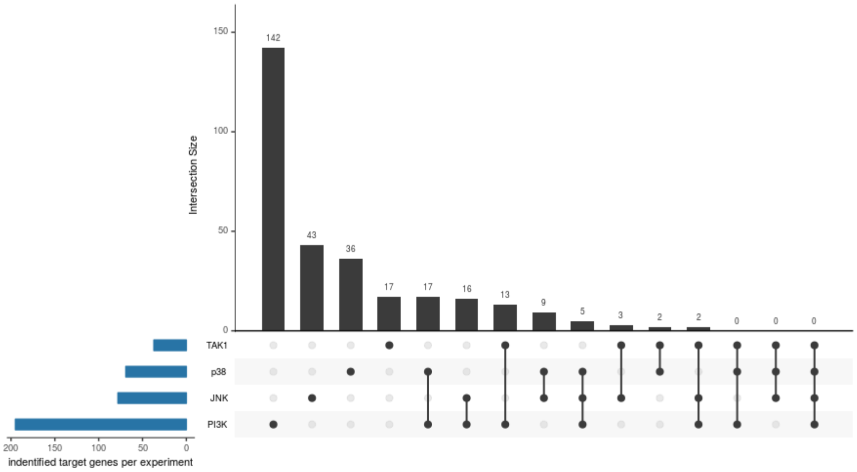


Figure 3.10: *Comparing the sets of consensus target genes of the different conditions. The size of the intersections is shown as a bar chart. Each column represents one intersection relationship. (142 were exclusively selected for condition PI3K; 42 target genes were only selected for condition JNK and so on.) Blue horizontal bars display the four set sizes. Plot generated by using the R package UpSetR from Conway et al. [2017].*

This way of visualization as in Figure 3.10 combines the set perspective and the element perspective. Related sets (here: intersections) are displayed as a matrix below the bar chart. The columns of this matrix correspond to the intersections while the rows correspond to sets, which means that each row matches a field in a Venn diagram. The number of elements in an intersection can be read from the length of the bars. That means the vertical histogram represents the size of the overlap between gene sets.

It can be seen that the most consensus genes were identified after PI3K inhibition (142 genes compared to 43 after inhibition of JNK, 36 after p38 inhibition, and 17 after TAK1 inhibition). This underlines the importance of PI3K in B cell lymphomas and has to be biologically validated. Also, it initiates the question with which pathways BCR signaling is intertwined. This can by the basis for following research, as the focus of this analysis lied more on BCR signaling in the context of feedback loops within the pathway.

3.2 Breast Cancer

3.2.1 Statistical analysis

In the differential analysis of the RNA-Seq and RPPA data, transcriptomic and proteomic profiles of the stimulated cell lines were compared to the control samples or other selected stimulation conditions.

3.2.1.1 RNA-Seq differential analysis

As the aim was to study the influence of the receptor ROR2 on WNT11 signaling, the main questions of the RNA-seq analysis were focused on: Which genes are regulated by WNT11

1. in the ROR2 overexpressing cells?

2. independently of ROR2 overexpression?

3. dependently of ROR2 overexpression?

Therefore, three cell lines with different characteristics were investigated individually. As described in section 2.5.2.1, one cell line had no endogenous ROR2, and two cell lines express endogenous ROR2. From these two ROR2 expressing cell, one cell line express endogenous WNT11 and in the second cell line WNT11 was silenced with siRNA. In that way, the observed effects could be directly associated with specific cellular conditions.

First of all, the RNA-Seq data set was first quality checked via FastQC (Babraham Bioinformatics) and then aligned to the transcriptome using STAR tool [*Dobin et al.*, 2013]. Gene-level abundances were estimated by RSEM algorithm [*Li and Dewey*, 2011]. Afterwards the R-Package edgeR [*Robinson et al.*, 2010] was used to perform the statistical analysis of the RNA-Seq data set, in order to identify differentially expressed genes were identified between different conditions by fitting negative binomial generalized linear models [*McCarthy et al.*, 2012]. Subsequently, Benjamini-Hochberg [*Hommel*, 1988] method was used to adjust the p values and significantly differentially expressed genes (DEG) were selected with a FDR level below 5%.

Subsequently, each cell line and each time point was investigated separately by comparing treated vs untreated condition individually. This means, that cells with and without recombinant human WNT11 (rhWNT) treatment were compared at all available time points. This should give a first impression of the underlying data before moving to the main research questions. Figure 3.11 shows the first 50 DEGs for the three cell lines separately.

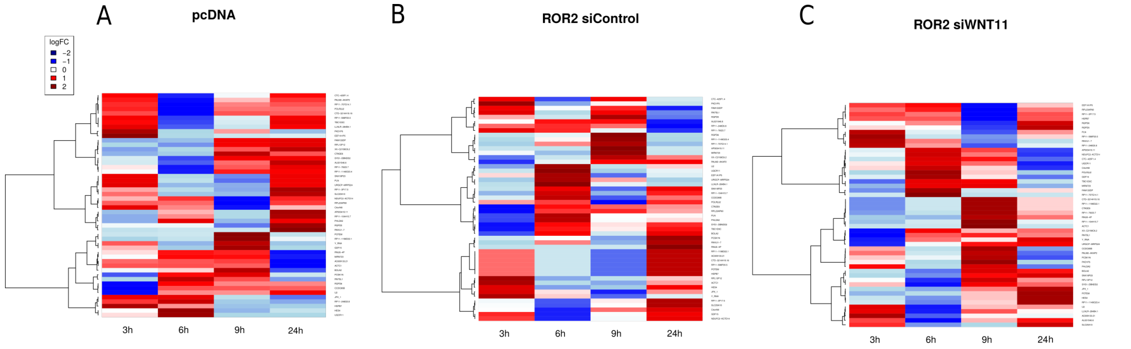


Figure 3.11: *Separate heatmaps for A: cells expressing the pcDNA empty vector B: ROR2 overexpression condition and control siRNA and C: ROR2 overexpression condition and siWNT11. All three heatmaps show the* $\log_2$ *FCs of phospho-proteins with dendrograms from hierarchical clustering analyses.*

The majority of these DEGs are non-protein coding genes. For this reason, the analysis was redone with an additional filtering step as described in section 2.1 keeping only those genes that have at least ten reads for some samples. From a methodical perspective, low counts do not provide enough statistical evidence to draw a reliable conclusion. Such genes, therefore, can be removed from the analysis without any loss of information. Furthermore, independent filtering can increase the sensitivity and the precision of DEGs [*Bourgon et al.*, 2010; *Chen et al.*, 2016].

In order to answer the aforementioned questions, the following comparisons were conducted after the pre-filtering step:

1. ROR2 overexpressing cells with and without siWNT11 (pROR2 siControl vs pROR2 + siWNT11) at time point 0h,
2. cells with empty vector with and without WNT11 (pcDNA vs. pcDNA + siWNT11) at time points 9h and 24h, and
3. ROR2 overexpressing cells without siWNT11 (pROR2 siControl) at time points 9h and 24h.

It is expected that stimulation effects on the gene expression level are more frequently observed at the later time points 9h and 24h. Following the central dogma of molecular biology, information transfer takes place in a sequential way from DNA to RNA and needs time to be apparent. Therefore, the focus was set on the two later time points.

The first comparison investigates the genes, which are regulated by endogenous WNT11 in ROR2 overexpressing cells. Indicating that the significant genes are regulated by endogenous WNT11 as these cells were not stimulated with rhWNT11. Among the significant DEGs (FDR< 0.05) some interesting candidates could be found. Among them are Ror2, Sonic hedgehog (Shh) and CD44. The protein encoded by Shh is a crucial gene involved in Hedgehog (HH) pathway. Even though the

potential role of SHH in breast cancer is not well described, recent studies showed its potential importance, particularly in aggressive triple-negative breast cancer subgroups. [*O'Toole et al.*, 2011]
The protein encoded by CD44 acts on the cell-surface and is involved in cell-cell interactions, and cell migration. CD44 is related with the WNT11 signaling pathway. [*Orian-Rousseau and Schmitt*, 2015]
These examples demonstrate that interesting effects could be observed and further investigations are necessary to validate the results and improve the understanding of the WNT11 signaling landscape. The complete list of significant DEGs resulting from this DEG analysis was later supplied to a pathway-based analysis with the approach called *pwOmics* (see section 2.2.3 for method description).

The lists of DEGs of the other two comparisons are part of separate investigations in order to follow the study of the effects of endogenous and external WNT11 stimulation in more detail.

3.2.1.2 RPPA differential expression analysis

Subsequently after preprocessing the raw data set as described in 2.1 differentially expressed phospho-proteins under particular conditions and time points were examined.

With attention to check which proteins are affected unspecifically by transfection, cells with ROR2 expression (pROR2) with ROR2 cells with empty transfection (pROR2 + siControl) were compared. This led to seven differentially expressed phospho-proteins.

To quantify the protein expression changes linked with silencing WNT11, the following comparison was performed: silenced WNT11 against not silenced WNT11 in cells with the empty vector (pcDNA vs. pROR2 + siWNT11) and cells with overexpression of ROR2 (pROR2 vs. pROR2 + siWNT11).

Heatmaps 3.12 across all time points representing the log_2 FC values between contrasted treatments for the 34 phospho-proteins for each time point. Phospho-proteins which are found to be differentially expressed (FDR < 0.05) in at least one time point are labeled in red. The value of log_2 FC between contrasted treatments for each gene is indicated by the colored scale, with red indicating up regulation and blue indicating down regulation.

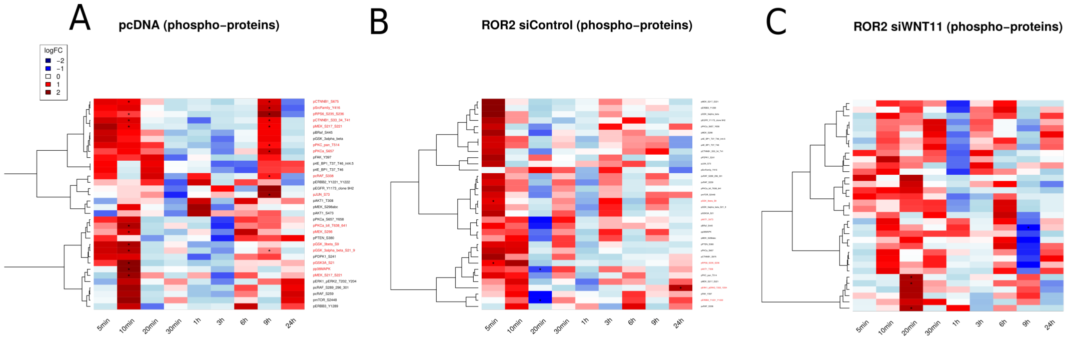


Figure 3.12: *Separate heatmaps for A: cells expressing the pcDNA empty vector B: ROR2 overexpression condition and control siRNA and C: ROR2 overexpression condition and siWNT11 showing* log_2 *FCs of phospho-proteins with dendrograms from hierarchical clustering analyses. Cells signed with "*" indicate a significant* log_2 *FC with FDR* < 0.05.

It can be seen in heatmap 3.12 A for cells expressing the pcDNA empty vector that the phosphorylation goes up very fast after 10 minutes and a second time after 9h.

Analog to the phospho-proteins, total proteins were analyzed with the same workflow. Figure 3.13 shows the log_2 fold-change values between contrasted treatments for the 67 total proteins for each time point.

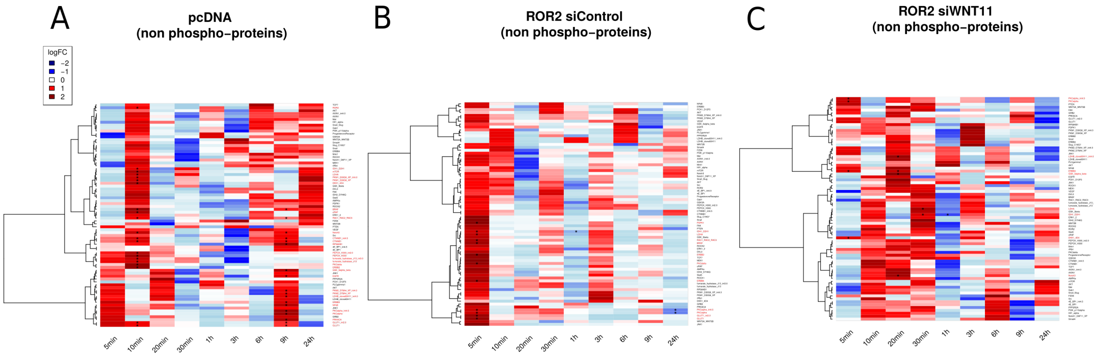


Figure 3.13: *Separate heatmaps for A: cells expressing the pcDNA empty vector B: ROR2 overexpression condition and control siRNA and C: ROR2 overexpression condition and siWNT11 showing* log_2 *FCs of total proteins with dendrograms from hierarchical clustering analyses. Cells signed with "*" indicate a significant* log_2 *FC with FDR* < 0.05.

3.2.2 DDEPN

The normalized time-course RPPA data are the basis for modeling the WNT11 signaling networks with the aim of exposing mechanisms of signal transduction dynamics in two of the cell lines. For network reconstruction, the method of DDEPN [*Bender et al.*, 2011] was applied to take advantage of the dynamical BN approach which is specially designed for longitudinal protein phosphorylation data measured in stimulation experiments.

As phosphorylation is a post-translational process, which alters the function of a protein through modifying the activity of an enzyme, network reconstruction was performed for the normalized phospho- and total proteins expression levels separately. To also benefit the most of this reconstruction method, it is advantageous to integrate background literature knowledge about protein interactions into the network model. Accordingly, before starting the main analysis, a sub-graph of the WNT11 pathways was manually compiled. 10 total proteins and 8 phospho-proteins were selected for the prior knowledge graphs. The two prior knowledge graphs, i.e. one for the total proteins and one for the phospho-proteins, are derived from the hsa04310 pathway from the KEGG database (http://www.genome.jp/kegg). The assembled graphs encompass a compilation of recent knowledge of the canonical and non-canonical WNT11 pathways.

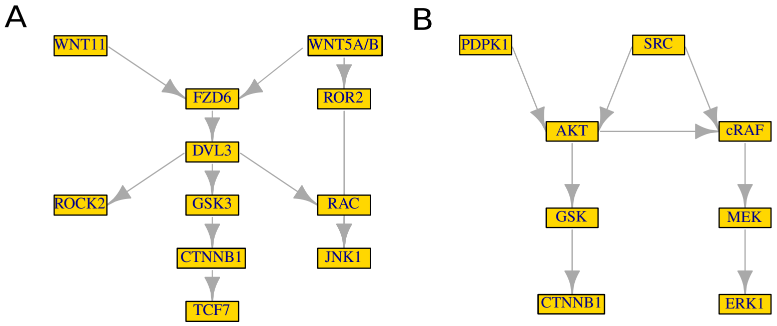


Figure 3.14: *A: Prior knowledge graph for total proteins. B: Prior knowledge graph for phospho-proteins.*

At first, reconstruction of the WNT11 signaling network for the total protein measurements was performed in both cell lines individually (Figure 3.15 A). It is observable, that the algorithm verified only the previously known edge between ROR2 → JNK1 in MCF-7 pcDNA cells and CTNNB1 → TCF7 in ROR2-overexpressing cells. Instead, different from the prior knowledge network, the analysis disclosed mostly (11 and 10) newly derived edges with more direct connections in pcDNA and ROR2-overexpressing cells.

In the results it is noticeable, that the algorithm could not exactly reproduce the signaling chain from FZD6 to TCF7. Instead, the method interpreted the observed effects as direct connections from FZD6 to GSK3, and CTNNB1 or from WNT11 to DVL3, GSK3 and TCF7. Comparing the signaling networks between the two cell lines, the algorithm found that WNT11 stimulation alters WNT5A/B and ROR2 expression independent of ROR2 overexpression. WNT5A/B is a representative ligand of the non-canonical WNT11 signaling pathway, who binds to FZD receptors together with different co-receptors, including ROR2. [*Verkaar and Zaman*, 2010] Its activation was observed in invasive breast cancer cells. [*Pukrop et al.*, 2006] On the other hand, WNT11 seems to directly alter CTNNB1, which is a central canonical WNT11 protein, only in ROR2-overexpressing cells. If CTNNB1 is not degraded, it accumulates in the nucleus. There, it acts as a transcriptional co-activator, initiating the cascade of downstream acting genes. [*Komiya and Habas*, 2008] The observation, that the edge ROR2 → CTNNB1 is only present in ROR2-overexpressing cells, indicates towards an influence of ROR2 receptor on downstream canonical WNT11 signaling. This is in line with a recent study, that identified an antagonistic function for ROR2 expression in regulating canonical WNT11 activity in vivo using a breast cancer mouse model. [*Roarty et al.*, 2017]

Secondly, reconstruction of the WNT11 signaling network for the phospho-protein measurements was conducted in both cell lines individually (Figure 3.15 B). A new, only two edges from the prior knowledge graph (GSK3B → CTNNB1, SRC → cRAF) were confirmed by the algorithm. Instead, the DDEPN algorithm estimated many new edges. In contrast to the total protein network, the two phospho-protein networks diverge significantly. They overlap only in two edges (WNT11 → PDPK1 and GSK3B → CTNNB1). The dissimilarity between the two cell lines might indicate that ROR2 overexpression has an impact on downstream WNT11 signaling. One central difference is the protein AKT. It is regulated via WNT11 and GSK3B only in the ROR2-overexpressing cells. The regulatory role of ROR2 on the activation of AKT was until now just shown in osteosarcoma cells. [*Dai et al.*, 2017] The results of this study indicate that ROR2 activates the AKT pathway and induces osteosarcoma cell migration. Also, it is known that the AKT pathway mediates many biological processes, such as proliferation, apoptosis, and growth. [*Vivanco and Sawyers*, 2002] The observation in the reconstructed networks indicates that there is an unsuspected connection between ROR2 expression and the AKT pathway, which should be biologically validated.

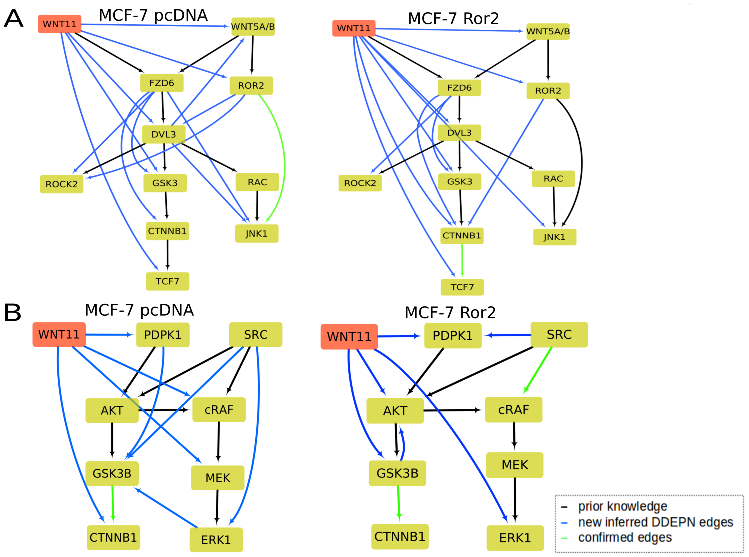


Figure 3.15: *WNT11 signaling network reconstruction of (A) total protein and (B) phospho-protein in MCF-7 pcDNA empty vector and ROR2-overexpressing cells. Yellow rectangles represent target proteins and red show the external stimulus WNT11.*

3.2.3 Survival Analysis

In order to gain a better insight into the role of selected WNT11 pathway members in the development of metastasis in breast cancer patients (see paragraph 2.5.2.2.2), was applied to pathways enrichment and survival analysis.

As part of Ms. Bayerlovás PhD thesis [*Bayerlová*, 2015], she examined several databases for pathways linked to WNT11 signaling. These were sorted into three subgroups reflecting different WNT11 signaling pathways: two gene sets with 304 and 489 genes representing the canonical and non-canonical WNT11 pathway as well as one set of genes with 173 genes acting upstream of the WNT11 pathways. [*Bayerlová et al.*, 2015]
The last set includes genes that deactivate or activate the WNT11 signals, as, for example, Hedgehog pathway members GLI genes [*He et al.*, 2006] and the MYC gene encoding the c-myc protein [*Cowling et al.*, 2007].

As a first step of this study, enrichment analysis was conducted on the basis of these three gene sets. The main purpose was to determine which of WNT11 signaling pathways is the most present pathway in breast cancer patients. Secondly, resulting clusters were provided to a Kaplan-Meier (KM) analysis of MFS. The KM curves were compared applying a log-rank test as implemented in *survival* R-package. Figure 3.16 A shows the enrichment of WNT11 pathways in the data set and the hierarchical clustering of the patients. Figure 3.16 B displays the MFS of the individual patient groups as Kaplan-Meier curves.

Albeit the different group sizes of the gene sets, a tendency can be identified in the results. As shown in the heatmap 3.16 A, the analysis revealed that genes involved in canonical WNT11 signaling are more expressed in primary breast cancer patients than the genes associated with the other two gene sets. Furthermore, the patients can be divided into three groups according to their WNT11 pathway activation levels. Focusing on the canonical WNT11 gene signature, hierarchical clustering shows that one patient cluster highly expresses canonical WNT11 pathway genes, the second cluster has a low expression level and the third cluster of patients can be classified as an intermediate patient group. Further, KM survival analysis revealed differences for MFS between the three groups ($p = 0.00421$, Figure 3.16 B).

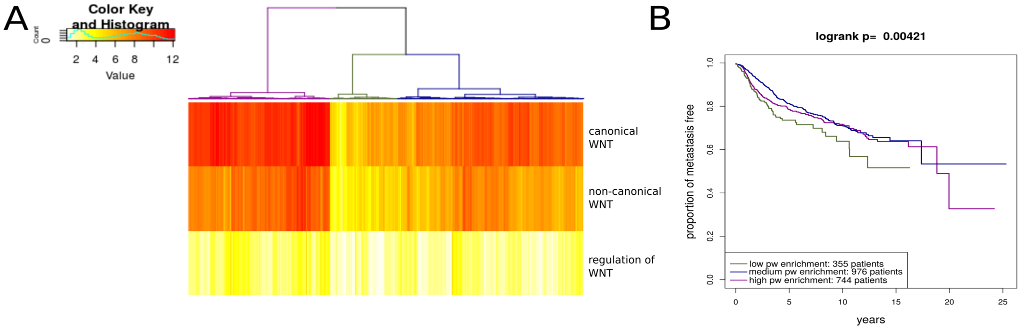


Figure 3.16: *The WNT11 gene signature sets in the patients cohort with primary breast cancer. A: Heatmap of $-log_2$ p values of log rank statistic, red color represents small p values and yellow color represents higher p values. Rows represent the three WNT11 signaling gene sets canonical, non-canonical pathway members and genes, that regulate WNT11. Cluster analyses yielded three patient clusters. B: Kaplan-Meier curves display metastasis free survival according to the three clusters.*

In primary breast cancer patients, it is noticeable that the patient groups differ in metastasis free years significantly. Consequently, the canonical WNT11 pathway is associated with the metastasis free survival.

Moving from primary breast cancer patients to measurements of brain metastasis of breast cancer patients, the same enrichment was performed on the publicly available data set of *Blazquez et al.* [2018]. The data set is described in paragraph 2.5.2.2.1. Figure 3.17 A shows a heatmap of 48 breast cancer patients with brain metastases.

As it can be seen in Figure 3.17 A, genes of the non-canonical WNT11 pathway is more present in patients with brain metastasis. Compared to the primary breast cancer patients, different WNT11 pathway cascades seem to be important for the development of metastasis. The hierarchical clustering yielded in two main patient clusters, which were supplied to KM analysis. The results reveal a central role for the genes of the non-canonical WNT11 pathway in patient survival. KM curves in Figure 3.17 B display that patients with high non-canonical WNT11 pathway enrichment have a shorter OS.

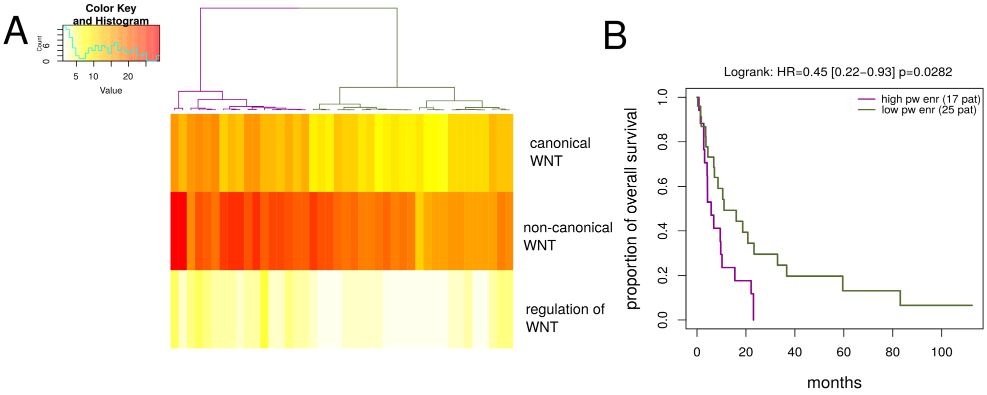


Figure 3.17: *The WNT11 gene signature sets in the patients cohort with brain metastasis. A: Heatmap of $-log_2$ p values of log rank statistic, red color represents small p values and yellow color represents higher p values. Rows represent the three WNT11 signaling gene sets canonical, non-canonical pathway members and genes, that regulate WNT11. Cluster analyses generated two patient clusters. B: Kaplan-Meier curves display OS according to the two clusters.*

The next attempt was to identify single gene markers from the WNT11 module that correlate with clinical outcome of the primary breast cancer patients. Therefore, the microarray data set of Ms. Bayerlová, again, was supplied to a third survival analysis. This time patients were grouped according to high and low expression levels of the genes WNT11, ROR2 as well as FZD4 and FZD6. For this, the median was chosen as reference expression value.

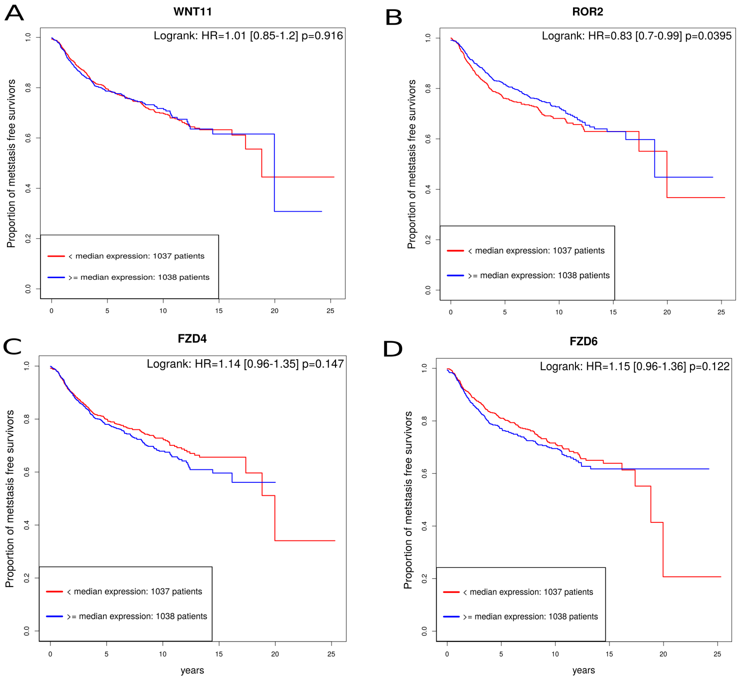


Figure 3.18: *Patients were grouped according to expression levels of (A) WNT11, (B) ROR2, (C) FZD4, and (D) FZD6 below or above the median expression level of each gene. Kaplan-Meier curves display metastasis free survival according to the two groups.*

As it can be seen in Figure 3.18, ROR2 (3.18 B) was detected at $p < 0.05$ level as significantly correlated with shorter MFS. Although the three other group splitting didn't yield in a significant result, the higher expression levels of the genes show a trend to result in shorter MFS. This indicates that these four genes have an impact on the development of metastasis of primary breast cancer patients. ROR2 and several FZDs have previously been described to be involved in the malignant transformation of breast cancer and to be associated with the survival and prognosis of patients. [*Bayerlová et al.*, 2017; *Zeng et al.*, 2018]To date, interplays between individual WNTs, FZDs and ROR2 remain poorly understood. Accordingly, further experimental studies are required to verify the findings of the present analysis and to understand the complexity of the underlying signaling mechanisms.

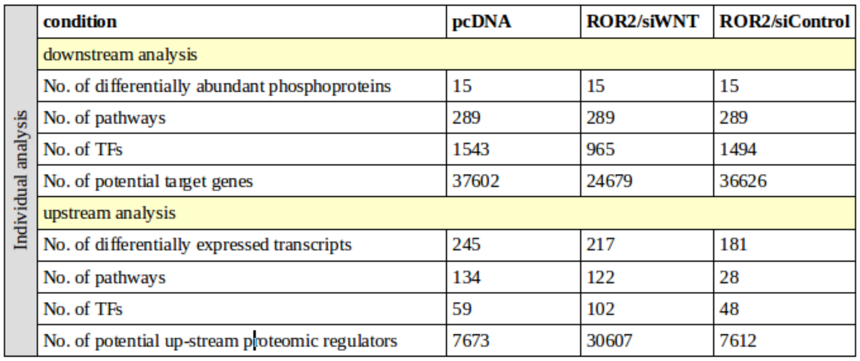
3.2.4 pwOmics

The next aim was to follow WNT11 dependent signaling from the activated receptors via proteins to the transcriptomic response. Like in the lymphoma project, the pathway-based integration approach, implemented in the R package *pwOmics*, was utilized for the comparative analysis of signaling structure on different cellular layers based on both data sets.

The workflow is basically the same as introduced in section 3.1.4. First, individual analysis of the two preprocessed omics data sets for each cell line was performed to consider these different functional layers. Therefore all measured phospho- and non phospho-proteins and significantly (FDR < 0.05) differentially expressed genes were selected for the down- and upstream analysis.

In the downstream analysis of the proteome data set, on the basis of pathway knowledge TFs of differentially abundant proteins and their target genes were identified. The upstream analysis of the transcriptome data set the TFs and proteomic regulators were determined based on differentially expression levels.

Table 3.3: *Result table of the individual analysis.*



	condition	pcDNA	ROR2/siWNT	ROR2/siControl
Individual analysis	downstream analysis			
	No. of differentially abundant phosphoproteins	15	15	15
	No. of pathways	289	289	289
	No. of TFs	1543	965	1494
	No. of potential target genes	37602	24679	36626
	upstream analysis			
	No. of differentially expressed transcripts	245	217	181
	No. of pathways	134	122	28
	No. of TFs	59	102	48
	No. of potential up-stream proteomic regulators	7673	30607	7612

Secondly, in the consensus analysis, consensus networks were reconstructed on the basis of joining proteins and genes from the previous down- and upstream analysis together with the corresponding PPI STRING network as described in section 2.2.3. This step resulted in an individual network for each cell line. Figures 3.19 A-C show the three received consensus networks after WNT11 stimulation.

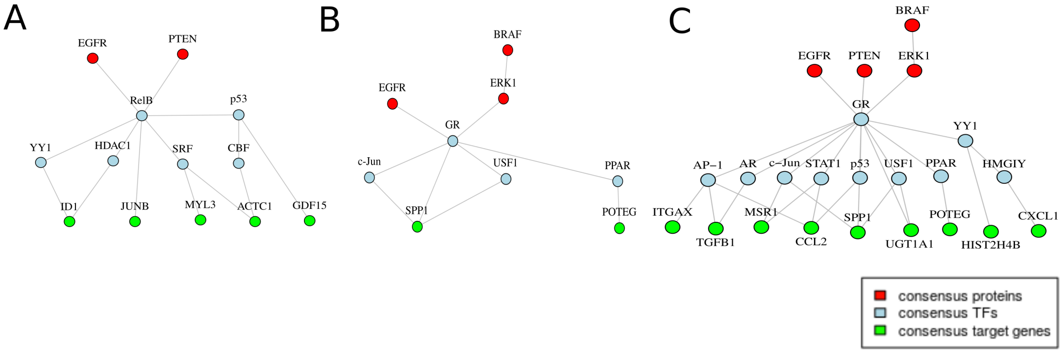


Figure 3.19: *Separate consensus graphs for A: cells expressing the pcDNA empty vector, B: cell with ROR2 overexpression and siWNT11 and C: cells with ROR2 overexpression control siRNA.*

The comparison of the graphs in Figure 3.19 shows that the consensus graph for ROR2 overexpessing cells transfected with siControl (pROR2 + siControl) (Figure 3.19 C) suggest as a subgraph of the consensus graph for ROR2 overexpessing cells transfected with siWNT11 (pROR2 + siWNT11) (Figure 3.19 B).

As a result, EGFR signaling was identified in all three cell lines as consensus-based regulatory process.
Interaction between WNT11 and EGFR has been identified in a few tumors. In breast cancers, WNT11 overexpression activates signaling via EGFR. [*Musgrove*, 2004] Several convergence points between the two pathways have been proposed, as reviewed in [*Hu and Li*, 2010]. Comparison of the consensus graphs in Figure 3.19 shows that ROR2 overexpression alters the downstream chains of EGFR. Thereby, fewer effects are observed after WNT11 stimulation in the cells with endogenous WNT11 (Figure 3.19 B) than in cells without endogenous WNT11 (Figure 3.19 C). The combination of ROR2 overexpression and presence or absence of endogenous WNT11 could provide an explanation for the observed differences.

The same pathway-based consensus analysis was applied to investigate signaling which is affected in ROR2 overexpressing cells by endogenous WNT11 signaling. Therefore, the significant DEGs and proteins resulting from the comparison of ROR2 overexpressing cells without endogenous WNT11 against ROR2 overexpressing cells with endogenous WNT11 were supplied to the *pwOmics* approach. The constructed consensus graph is shown in Figure 3.20. Due to the large number of DEGs (3789 genes with FDR < 0.05) which were supplied to the algorithm, a rather big consensus graph was derived.

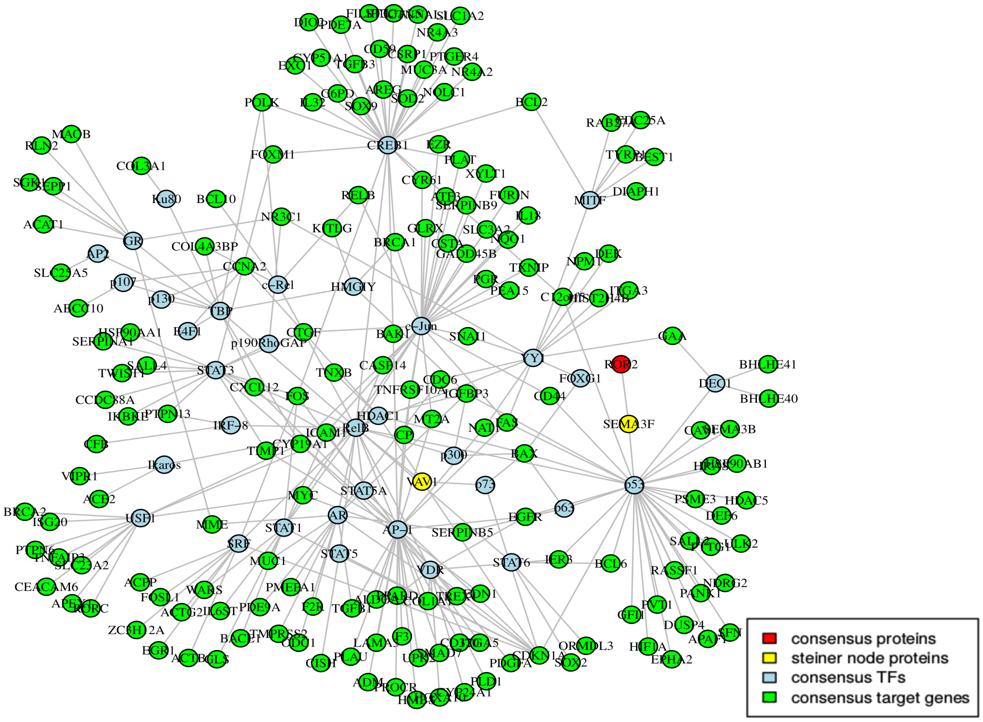


Figure 3.20: *Consensus graph for ROR2 overexpressing cells at time point 0h. Consensus protein nodes are colored in red, consensus target gene nodes in green, consensus transcription factor nodes in blue and Steiner nodes ind yellow.*

Among the identified consensus genes are pathway members of, for instance, PI3K/AKT signaling pathway and JAK/STAT signaling pathway.

In a subsequent step, literature was examined in order to determine what is known about these pathways. With it, evidence could be found that these signaling pathways have an important role in breast cancer. For instance, in triple-negative breast cancer, oncogenic activation of the PI3K/AKT pathway can by its overexpression of one of its upstream regulator EGFR. [*Kallergi et al.*, 2008; *Costa et al.*, 2018] This is consistent with the preceding analysis in which EGFR signaling was identified as an altered pathway in ROR2 overexpressing cells.
Likewise, a relationship between JAK/STAT activation and prognosis of breast cancer patients has been observed. Lack of activated STAT5 is associated with decreased survival and drives tumor progression and metastasis. [*Peck et al.*, 2011; *Banerjee and Resat*, 2016; *Chang et al.*, 2013]

CHAPTER IV

Discussion and Outlook

4.1 Discussion

Cancer biology is an astonishing and dynamic field of research. It has the general aim to determine the factors that make up a living system like a cell and to understand the interactions between them that result in the (mal-)functioning of the system. In the recent past, cancer research has become more and more driven by high-throughput technologies to generate biological data. As a result of the enormous amount of data, cancer biology research needs computational and statistical methods to analyze the data. Therefore, in genomics and systems biology applications the attention was moved towards identifying essential 'functional outputs'. In particular, as it was started to understand that cancer is not a disease of genes but of pathways.

Generally, pathways work by protein signaling transduction events that in the end drive changes in gene transcription.[*Hanahan and Weinberg*, 2000] Post-translational alterations may not be observed at the gene transcription level. Therefore gene transcription data provide information on the downstream effects of deregulated signaling. On the other hand, pathways emerging upstream (of an observed transcriptional pattern) could better be verified employing proteome data.

Over the last few years, the development of new measurement platforms to globally profile the cell at various molecular levels, including (among others) mRNAs and proteins, provides us challenges and capabilities to integrate these individual data types in relevant ways.
In recent years, the potential to study cellular and molecular systems has been revolutionized as a result of the expansion of omics sciences. An almost unlimited power lies within huge omics data sets. The success of omics sciences to further our understanding of human disease remains difficult because of several factors. One key reason postulated is that while individual omics domains yield distinct and important information, no single omics science is sufficient to facilitate a comprehensive understanding of the complex human biology and physiology. Therefore, multi-omics data, e.g. multiple types of biological data, should be considered to describe such complex biological processes.

As biological and medical scientists are interested in integrating recently measured data with earlier published clinical and prior biological results. Furthermore, data have been produced in different formats (graphs, sequences, etc.) and dimensions. The integrative approaches attempt to investigate an adequately large amount of samples and to cover the numerous sources of variability in their statistical models. Thus, omics studies benefit from the collection of large data sets.

In this work, newly generated data sets and publicly available data sets were investigated to identify signaling network structures in cancer cells. In two separately conducted studies, different network approaches were used to validate exiting knowledge and to identify new interactions between genes or proteins.

4.1.1 Lymphomas

The awareness that the survival or proliferation of B cell lymphomas depends on their interaction with the micro-environment, as well as on the expression of the B cell receptor *Kraus et al.* [2004]; *Lam et al.* [1997] might contribute to novel treatment strategies. The main question within this project was declared: How can current knowledge about B cell lymphomas be improved by combing existing methods? To answer this question, first, separate networks for gene and phospho-protein expression data were generated.

Based on the above, it can be concluded that both applied methodologies can be of use to analyze the data sets, but are in our case not sufficient to integrate the phospho-protein and gene expression measurements. To bring both data sets together a bigger agreement between measured and inhibited proteins with perturbation experiments for the micro-array data is needed. Nevertheless, to integrate both data sets and to overcome the restraints are derived from the small overlap, existing knowledge collected from literature databases was included subsequently. Thus, a method was applied, which aims to integrate two data sets using literature information. In this way, the analysis could benefit from biological pathway knowledge.

To ensure better integration and comparison of the resulting graphs, it is necessary to have a higher overlap between performed perturbation experiments of both measurements and also with the phospho-proteins, which are quantified. A wider overlap between nodes of graphs, which are coming from different network approaches is required to enable to match and integrate both results. In this case, the nodes of the DDEPN graph represent the measured phospho-proteins and the nodes of the NEM graph are interpreted as inhibitors. For that reason, it was not revealing to combine the data sets without the integration of a literature-based analysis step. There, consensus graphs for each treatment condition were produced.
To overcome this, extensive integration of literature databases and to learn directly from the underlying data sets, a larger overlap between the measured phospho-proteins and the performed inhibition experiments in the proteomic and transcriptomic

measurements are required. For further research, it is advantageous to prepare data sets with the same experimental setting in order to improve information about interaction patterns and to understand underlying biological processes.

All of the applied methods could not identify feedback loops that are reported in recent biological studies. This is because feedback loops cannot be easily identified by these classical network methods, which only determine signaling flow in a linear pathway but not within a feedback loop, where the participating elements are upstream and downstream at the same time. The relationship between p38 MAPK and the MEK/ERK signaling pathway is so far not fully understood. Different studies indicate an upstream influence of p38 MAPK on the MEK/ERK pathway. For instance, *Hirosawa et al.* [2009] demonstrated an increase of ERK1/2 phosphorylation after p38 MAPK inhibition, which suggests and negative influence of p38 MAPK on ERK1/2. On the other side, *Chen et al.* [2000] detected a direct negative feedback mechanism on RAS activity by p38 MAPK signaling. Particularly regarding consequences for therapeutic response and avoiding the development of drug resistance, (therapeutic) agents targeting, for instance, the ERK1/2 pathway should also inhibit the relevant (negative) feedback loops. Consequently, it is important to understand the pathway mechanisms and study the existence of feedback loops in lymphoma cells.

Modeling challenges faced in the study of lymphomas included the merging of data with missing experiments or measurements and also the need for individual data analysis expertise. Some limitations could be addressed by an extended inclusion of literature databases. Utilizing the integration approach *pwOmics* [*Wachter and Beißbarth*, 2015] enabled the identification of interesting pathway interplays downstream of BCR. These results provide the basis for biological validation and further investigations.

Furthermore, network reconstruction requires a good balance between prior knowledge and data driven modeling. Network inference must have three key features:

(i) The method should only consider the part of the prior which supports the data. This is necessary because the prior information usually is a set of possible interactions, of which just a subset might be relevant. In addition, this involves robustness to false interactions in the prior, coming from different sources.

(ii) Using a prior should not restrict the capability to discover the part of the network for which no prior information exists.

(iii) The user should be able to control the power given to the prior to adjust the method based on the trust in the prior.

Principally, the approach DDEPN, used to analyze the phospho-proteome data, allows to control the prior influence with a hyperparameter, which determines the weight of the prior knowledge during network reconstruction.

As already mentioned, NEM and DDEPN are designed to model the way of perturbation effect propagation in the networks from parent to child nodes. Therefore they are not able to simulate feedback loops, which were, in this case, expected from the biological data. If the focus lies more in the study of feedback loops in a pathway, the considered mathematical models should be able to estimate multiple interactions and reproduce existing evidence. To get more insights into feedback loops, it can beneficial to incorporate more dynamical approaches like, for instance, ordinary differential equations (ODE). They have already been used to investigate various biological behaviors such as the dynamics of positive and negative feedback loops. [*Shin et al.*, 2009] Statistical simulations of these methods allow identifying the strength of hypothesized causal pathways based on observed data.

4.1.2 Breast Cancer

Being breast cancer one of the most common cancer type among women, there are enormous studies carried out for breast cancer treatment. But the outcome seems to be insufficient and demands further extensive research, which could unveil the specific targets for breast cancer.

The focus of this project was the role of WNT ligands, especially ROR2 on WNT11 signaling in breast cancer.

The dysregulation of WNT11 signaling is directly associated with cancer. The increased development of sequencing technologies allows us to outline the heterogeneous molecular characteristics of breast cancer cells. WNT11 signaling is highly complex and not yet fully characterized. The discovery of novel regulators, such as ROR2, adds to the complexity but also presents exciting new opportunities for the development of potential therapeutic targets.
As the ROR2 overexpression targets have been implicated in the invasiveness of breast cancer cells, the association of the non-canonical WNT11 pathway members and metastasis free years of primary breast cancer patients was investigated.

In this study, several methods and data sets were combined to study the role of WNT11 signaling pathways in breast cancer. Therefore, gene and protein expression measurements from breast cancer cell lines were explored as well as survival analysis of patient data. The methods utilized range from differential expression analysis over network reconstruction to survival analysis.

To investigate the WNT11 signaling network influenced by ROR2 overexpression, network reconstruction based on RPPA measurements was conducted. The networks for total proteins and phospho-proteins were evaluated separately. To investigate the influence on the WNT11 signaling, the pathways were constructed for ROR2-positive and -negative cells individually. The comparison of the resulting pathways in the two cell lines revealed that the utilized network approach, DDEPN, focuses on the exploration of WNT11 signaling in cell lines with and without ROR2 overexpression.

Therefore, networks of phospho- and total proteins were studied separately. The phospho-protein networks displayed an interaction between ROR2 and CTNNB1 in ROR2 overexpressing cells. CTNNB1 serves as a co-activator in the canonical WNT11 signaling pathway. [*Cara et al.*, 2012] This indicates an influence of ROR2 on downstream canonical WNT11 signaling. To study potential receptors of the canonical WNT11 pathway is important for the discovery of new drug targets. It is known that the canonical WNT11 signaling pathway is highly activated in malignant breast cell lines. [*Benhaj et al.*, 2006]

The results of total protein networks also revealed an interesting link to the AKT pathway, indicating that it might become activated in breast cancer cells by diverse mechanisms. Given the important role of AKT signaling in regulating processes such as cell growth, proliferation, and survival, it is comprehensible that components of this pathway are (dys)regulated in cancer. The understanding of the complex interaction between members of intertwined pathways is necessary for the development of anti-cancer drugs that are targeted against this pathway. Some of the results could be published in [*Sitte et al.*, 2019].

After analyzing the interplay of selected pathway members, the further investigations focused on linking the different WNT11 pathways and the expression of some of their members with clinical outcomes. Therefore, genes, that are associated with WNT11 signaling were divided into three groups. Ms. Bayerlová described in her PhD thesis [*Bayerlová*, 2015] three gene sets. Two of them represent the canonical and the non-canonical WNT11 pathway. The third gene set collects all genes that are known as upstream regulators of the WNT11 pathway. To learn more about the WNT11 signaling in breast cancer in a clinical context, microarray data of breast cancer patients was included in the analysis. The first data set was a microarray data set, which is a compilation of different publicly available breast cancer patient data, collected by [*Bayerlová*, 2015]. The second was RNA-seq measurements of breast cancer patients with brain metastasis. Based on the three gene sets, enrichment analysis was applied on both patient data sets, to further reveal the role of the different WNT11 pathways. This yielded an important role of non-canonical WNT11 pathway members in the development of metastasis, and, in contrast, an important role of genes known as canonical WNT11 pathway members in the overall survival of breast cancer patients with brain metastasis.

In a further attempt, four single gene markers from the WNT11 pathway were tested if their expression is correlated with metastasis outcome in breast cancer patients. With this, it could be shown that patients with higher ROR2 expression develop earlier metastasis than patients with lower ROR2 expression ($p < 0.05$), but there is no significant difference between patients with high and low WNT11 expression. This suggests altered downstream signaling initiated by ROR2. This is in line with recent studies who could associate ROR2 expression with metastasis and tumor progression.

[*Bayerlová et al.*, 2017; *Roarty et al.*, 2017]

Finally, to jointly analyze the transcriptomic and proteomic data sets, pathways databases were investigated. The aim was to identify consensus molecules that are significantly regulated in both data sets. The results point towards the relationship between ROR2 overexpression and an altered EGFR signaling in the breast cancer cells. In the last couple of years, it has been studied that both, WNT and EGFR signaling, are closely related to tumorgenesis. Recent studies found evidence that WNT11 and EGFR crosstalk with each other in cancer development. [*Hu and Li*, 2010; *Schlange et al.*, 2007] EGFR mediated PI3K/AKT activation stimulates β-catenin transactivation and increases tumor cell invasion. This indicates that EGFR activation transactivates β-catenin via WNT11 signaling pathways in tumor cells. [*Sharma et al.*, 2002; *Zhimin et al.*, 2003]
Reassuringly, AKT signaling was also identified on phospho-protein level as an altered pathway affected by ROR2 overexpression and external WNT11 stimulation. Previous research could already investigate aberrant AKT expression as an important signaling hub in cancer cells. [*Sharma et al.*, 2002; *Vivanco and Sawyers*, 2002; *Costa et al.*, 2018] Also, this illustrates a more complex WNT11 signaling topology and different pathways are intertwined. Changes in upstream signaling molecules will not simply alter WNT11 signaling, but also other parallel pathways that are regulated by these upstream activators. Therefore, reinforcing studies are needed to unravel the whole network machinery.

In spite of the aforementioned results, there are still some open questions in this study. The predicted results, such as expression of the examined genes and proteins, and the relations between them and WNT11 signaling pathway, are required to be confirmed by experiments in cancer tissues.

In conclusion, on one side, this work demonstrates how a systematic approach can be applied for the identification of the relationship between WNT11 receptors and downstream signaling routes. On the other side, the results together might suggest that the role of WNT11 signaling pathway is more complex and that different WNT11 receptors function in different ways.

4.2 General Conclusion and Outlook

This thesis brings open bioinformatics tools together in one workflow that allows to analyze coupled transcriptome and proteome measurements. Overall, this work has strengths and limitations that should be mentioned. This study investigates various ways of network reconstruction, namely NEMs and DDEPNs, that can be inferred from gene and protein expression data sets and the advantage of incorporating biological knowledge in such methods. Bioinformatic approaches adapted from graph theory concepts as well as enrichment analysis were shown to be suitable and strong instruments to break down complex protein expression patterns, as in the reconstruction of WNT11 signaling in the breast cancer cell line MCF-7. In this context, the applied network analysis approach DDEPN identified interesting interplays downstream of WNT11 and ROR2 after external WNT11 stimulation.

The work is as a methodological contribution to the analysis of studies with different data sets with a considerable overlap of samples.

Hence, the experimental design and data set selection are important factors to consider for data analysis and consequently for any bioinformatical integrative investigation. For example, the analysis approach illustrated in the lymphoma study assumes that gene or protein expression samples have been extracted under two or more treatment conditions. Therefore, this approach is especially appropriate for experiments with stimulation or perturbation designs.
In contrast, the breast cancer study comprised more data sets, which made it possible to apply, additionally to network approaches, survival analysis techniques to associate the findings from network reconstruction with clinical outcome.

All things considered, bioinformatic approaches adapted from graph theory concepts as well as enrichment analysis were shown to be suitable and strong instruments to break down complex gene expression patterns.
Discovering relevant links between important signaling molecules in a pathway and their predictive outcomes is the key to discover new drug targets. The ability to produce a detailed characterization of a disease allows the stratification of patients into well-defined groups for tailored treatment.

Confounding problems made data integration a not trivial task. The types of data to be integrated range from smaller assays with a selected list of protein expression levels to high-throughput mRNA measurements. Whereby each technology used brings its own different degree of reliability. In addition to data generated by high-throughput technologies, other sources of data, such as clinical data and curated databases, can be utilized. Curated databases may also combine data from different experimental conditions. Computational calculations that utilize them risk to continue underlying systematic bias. These concerns have to be addressed in every integrative analysis.

The integration workflow presented here showed some limitations that need to be addressed. Combining two different omics data sets requires an appropriate experiment setting in order to have the same conditions or time points for both data sets. The analysis of the Lymphoma data sets was limited by missing overlapping conditions. This is due to the layout of the study, which attempts to combine data from different individual projects.
Some of the limitations can be compensated by integrating new data sets with prior knowledge, but the perfect advantage of integrated analysis can be obtained only if data acquisition from all utilized platforms are designed equally by multidisciplinary teams. As demonstrated in the breast cancer study, the beforehand designed outline will allow a better comparison and integration of the transcriptome and proteome

results.

An important consideration in all bioinformatical studies is how to validate results both at the biological functional level and at the replication level. The presented results raise interesting hypotheses, which must be confirmed by a follow-up experimental validation. In this thesis, a biological validation is ongoing work and not finished at the moment of submitting the thesis.

This thesis demonstrates on the one hand that it is still challenging to analyze different data sets in a combined way, but on the other hand, it illustrates that the integration of multiple data sets allows the identification of new information, which could not identified when only one data set was considered.

The development of statistical methods aiming at the integration of more than one single data set will be crucial to obtain information on complex diseases such as cancer. The presented approaches are potential tools that can be used comprehensively in the study of the signaling networks in different complex diseases that can ultimately lead to the discovery of relevant links between important signaling molecules in a pathway and their predictive outcomes, which is the key to discover new drug targets. The ability to produce a detailed characterization of a disease allows the stratification of patients into well-defined groups for tailored treatment.

Bibliography

Adams, MD, et al. (1991), Complementary dna sequencing: expressed sequence tags and human genome project, *Science*, *252*(5013), 1651–1656.

Alberts, B, Johnson, A, Lewis, J, Raff, M, Roberts, K, , and Walter, P (2007), *Molecular Biology of the Cell*, 5 ed., Garland Science, New York.

Aldoss, IT, Weisenburger, DD, Fu, K, Chan, WC, Vose, JM, Bierman, PJ, Bociek, RG, , and Armitage, JO (2008), Adult burkitt lymphoma: advances in diagnosis and treatment, *Oncology*, *22*, 1508–1517.

Alvarez-Chaver, P, Otero-Estevez, O, Paez de la Cadena, M, Rodriguez-Berrocal, FJ, and Martinez-Zorzano, VS (2014), Proteomics for discovery of candidate colorectal cancer biomarkers, *World Journal of Gastroenterology*, *20*(14), 38043824.

Anders, S, and Huber, W (2010), Differential expression analysis for sequence count data, *Genome Biology*, *11*(10), R106.

Andrews, S (2010), Fastqc: a quality control tool for high throughput sequence data, `http://www.bioinformatics.babraham.ac.uk/projects/fastqc`, online; accessed 2020-02-10.

Azeloglu, EU, and Iyengar, R (2015), Signaling networks: Information flow, computation, and decision making, *Cold Spring Harbor Perspectives in Biology*, *7*(4).

Bader, GD, Cary, MP, and Sander, C (2006), Pathguide: a pathway resource list, *Nucleic Acids Research*, *34*, D504–D506.

Balaji, S, Ahmed, M, Lorence, E, Yan, F, Nomie, K, and Wang, M (2018), Nf-κb signaling and its relevance to the treatment of mantle cell lymphoma, *J of Hematology & Oncology*, *11*(1), 83.

Balbin, OA, et al. (2013), Reconstructing targetable pathways in lung cancer by integrating diverse omics data, *Nat Commun*, *4*, 2617.

Banerjee, K, and Resat, H (2016), Constitutive activation of stat3 in breast cancer cells: A review, *International Journal of Cancer*, *138*(11), 2570–2578.

Barbulovic-Nad, I, Lucente, M, Sun, Y, Zhang, M, Wheeler, AR, and Bussmann, M (2006), Bio-microarray fabrication techniques - a review, *Critical Reviews in Biotechnology*, *4*(26), 237–259.

Batushansky, A, Toubiana, D, and Fai, A (2016), Correlation-based network generation, visualization, and analysis as a powerful tool in biological studies: A case study in cancer cell metabolism, *BioMed Research International*, *2016*.

Bayerlová, M (2015), Pathway and network analyses in context of wnt signaling in breast cancer, Ph.D. thesis, Georg-August-University Göttingen.

Bayerlová, M, Klemm, F, Kramer, F, Pukrop, T, Beißbarth, T, and Bleckmann, A (2015), Newly constructed network models of different wnt signaling cascades applied to breast cancer expression data, *PLoS One*, *10*(12), e0144,014.

Bayerlová, M, Menck, K, Klemm, F, Wolff, A, Pukrop, T, Binder, C, Beißbarth, T, and Bleckmann, A (2017), Ror2 signaling and its relevance in breast cancer progression, *Frontiers in Oncology*, *7*, 135.

Bellacosa, A, Chan, TO, Ahmed, NN, Datta, K, Malstrom, S, Stokoe, D, McCormick, F, Feng, J, and Tsichlis, P (1998), Akt activation by growth factors is a multiple-step process: the role of the ph domain, *Oncogene*, *17*, 313325.

Bender, C, Henjes, F, Fröhlich, H, Wiemann, S, Korf, U, and Beißbarth, T (2010), Dynamic deterministic effects propagation networks: learning signalling pathways from longitudinal protein array data, *Bioinformatics*, *362*, 365–386.

Bender, C, von der Heyde, S, Henjes, F, Wiemann, S, Korf, U, and Beißbarth, T (2011), Inferring signalling networks from longitudinal data using sampling based approaches in the r-package 'ddepn', *Bioinformatics*, *12*, 291.

Benhaj, K, Akcali, KC, and Ozturk, M (2006), Redundant expression of canonical wnt ligands in human breast cancer cell lines, *Journal of the Royal Statistical Society*, *15*, 701707.

Benjamini, Y, and Hochberg, Y (1995), Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society*, *57*(21), 289300.

Bio-Plex (1999), Bio-plex® multiplex immunoassays, `https://www.bio-rad.com/de-de/applications-technologies/bio-plex-multiplex-immunoassays?ID=LUSMOZMNI`, online; accessed 2020-02-10.

Blazquez, R, et al. (2018), Pi3k: A master regulator of brain metastasis-promoting macrophages/microglia, *Glia*, *66*(11), 2438–2455.

Bojarczuk, K, et al. (2019), Targeted inhibition of pi3kα/δ is synergistic with bcl-2 blockade in genetically defined subtypes of dlbcl, *Blood*, *133*(1), 70–80.

Bourgon, R, Gentleman, R, and Huber, W (2010), Independent filtering increases detection power for high-throughput experiments, *Proceedings of the National Academy of Sciences*, *107*(21), 9546–9551.

Bradford, JR, Needham, CJ, Bulpitt, AJ, and Westhead, DR (2006), Insights into protein-protein interfaces using a bayesian network prediction method, *Journal of Mol Biol*, *338*, 175–190.

Burkitt, D (1969), Etiology of burkitt's lymphomaan alternative hypothesis to a vectored virus, *J Natl Cancer Inst*, *42*, 19–28.

Campo, E, Swerdlow, SH, Harris, NL, Pileri, S, Stein, H, and Jaffe, ES (2011), The 2008 who classification of lymphoid neoplasms and beyond: evolving concepts and practical applications, *Blood*.

Cara, J, Manisha, S, and Beric, RH (2012), Wnt signaling from membrane to nucleus: β-catenin caught in a loop, *The International Journal of Biochemistry and Cell Biology*, *44*(6), 847–850.

Chang, Q, et al. (2013), The il-6/jak/stat3 feed-forward loop drives tumorigenesis and metastasis, *Neoplasia*, *15*(7), 848–862.

Chen, G, Hitomi, M, Han, J, and Stacey, DW (2000), The p38 pathway provides negative feedback for ras proliferative signaling, *The Journal of Biological Chemistry*, *275*(50), 38,973–38,980.

Chen, R, et al. (2012), Personal omics profiling reveals dynamic molecular and medical phenotypes, *Cell*, *148*(6), 1293–1307.

Chen, Y, Lun, ATL, and Smyth, GK (2016), From reads to genes to pathways: differential expression analysis of rna-seq experiments using rsubread and the edger quasi-likelihood pipeline, *F1000Research Rev*, *5*(1438).

Chu, DH, Morita, CT, and Weiss, A (1998), The syk family of protein tyrosine kinases in t-cell activation and development, *Immunol Rev*, *165*, 167–180.

Chuang, H-Y, Lee, E, Liu, Y-T, Lee, D, and Ideker, T (2007), Network-based classification of breast cancer metastasis, *Molecular Systems Biology*, *3*(1).

Ciriello, G, Cerami, E, Sander, C, and Schultz, N (2012), Mutual exclusivity analysis identifies oncogenic network modules, *Genome Res*, *22*(2), 398406.

Collins, MO, Yu, L, and Choudhary, JS (2007), Analysis of protein phosphorylation on a proteomescale, *Proteomics*, *7*(16), 95106.

Conway, JR, Lex, A, and Gehlenborg, N (2017), Upsetr: an r package for the visualization of intersecting sets and their properties, *Bioinformatics*, *33*(18), 2938–2940.

Costa, BM, et al. (2011), Impact of egfr genetic variants on glioma risk and patient outcome, *Cancer Epidemiology and Prevention Biomarkers*, *20*(12), 2610–2617.

Costa, RLB, Han, HS, and Gradishar, WJ (2018), Targeting the pi3k/akt/mtor pathway in triple-negative breast cancer: a review, *Breast Cancer Research and Treatment*, *169*(3), 397–406.

Cowling, VH, D'Cruz, CM, Chodosh, LA, and Cole, MD (2007), c-myc transforms human mammary epithelial cells through repression of the wnt inhibitors dkk1 and sfrp1, *Molecular and Cellular Biology*, *27*(14), 5135–5146.

Crespo, M, et al. (2006), Zap-70 expression in normal pro/pre b cells, mature b cells, and in b-cell acute lymphoblastic leukemia, *Clin.Cancer Res*, *12*, 726734.

Croft, D, et al. (2014), The reactome pathway knowledgebase, *Nucleic Acids Res*, *42*, D472–7.

Csardi, G, and Nepusz, T (2006), The igraph software package for complex network research, *InterJournal*, *Complex Systems*, 1695.

Dai, B, Yan, T, and Zhang, A (2017), Ror2 receptor promotes the migration of osteosarcoma cells in response to wnt5a, *Cancer Cell Int*, *17*, 112.

Dal Porto, J, Gauld, S, Merrell, K, Mills, D, Pugh-Bernard, A, Cambier, J, and Cambier, J (2004), B cell antigen receptor signaling 101, *Mol. Immunol*, *41*, 599–613.

D'Ambrosio, D, Cantrell, DA, Frati, L, Santoni, A, and Testi, R (1994), Involvement of p21ras activation in t cell cd69 expression, *Eur. J. Immunol*, *24*, 616–620.

David, H (1999), A tutorial on learning with bayesian networks, in *Learning in Graphical Models*, edited by MI, Jordan, pp. 301–54, MIT Press, Cambridge, MA, USA.

De, A (2011), Wnt/ca^{2+} signaling pathway: a brief overview, *Acta Biochimica et Biophysica Sinica*, *43*(10), 745–756.

de la Fuente, A, Bing, N, Hoeschele, I, and Mendes, P (2004), Discovery of meaningful associations in genomic data using partial correlation coefficients, *Bioinformatics*, *20*(18), 3565–3574.

de Lorenzo, V (2008), Systems biology approaches to bioremediation, *Curr. Opin. Biotechnol*, *19*, 579589.

DeFranco, A (1997), The complexity of signaling pathways activated by the bcr, *Curr. Opin. Immunol*, *9*, 296–308.

Dettmer, K, Aronov, PA, and Hammock, BD (2007), Mass spectrometry-based metabolomics, *Mass Spectrom Rev*, *29*(1), 51–78.

Dhillon, A, Hagan, S, Rath, O, and Kolch, W (2007), Map kinase signaling pathways in cancer, *Oncogene*, *26*, 3279–90.

Dobin, A, Davis, CA, Schlesinger, F, Drenkow, J, Zaleski, C, Jha, S, Batut, P, Chaisson, M, and Gingeras, TR (2013), Star: ultrafast universal rna-seq aligner, *Bioinformatics*, *29*(1), 15–21.

Dougherty, MK, et al. (2005), Regulation of raf-1 by direct feedback phosphorylation, *Molecular Cell*, *17*(2), 215–224.

Doğrusöz, U, Madden, B, and Madden, P (1997), Circular layout in the graph layout toolkit, in *Graph Drawing*, edited by North, S, pp. 92–100, Springer Berlin Heidelberg, Berlin, Heidelberg.

Dutkowski, J, Kramer, M, Surma, MA, Balakrishnan, R, Cherry, JM, Krogan, NJ, and Ideker, T (2013), A gene ontology inferred from molecular networks, *Nat Biotechnol*, *1*(31), 38–45.

Eduati, F, De Las Rivas, J, Di Camillo, B, Toffolo, G, and Saez-Rodriguez, J (2012), Integrating literature-constrained and data-driven inference of signalling networks, *Bioinformatics*, *28*(18), 23112317.

Ellson, J, Gansner, E, Koutsofios, L, North, SC, and Woodhull, G (2002), Graphviz - open source graph drawing tools, in *Graph Drawing*, edited by Mutzel, P, Jünger, M, and Leipert, S, pp. 483–484, Springer Berlin Heidelberg, Berlin, Heidelberg.

Emmert-Streib, F, and Dehmer, M (2011), Networks for systems biology: conceptual connection of data and function, *IET Syst Biol*, *5*(3), 185–207.

Engel, P, Zhou, L, Ord, D, Sato, S, Koller, B, and Tedder, T (1995), Abnormal b lymphocyte development, activation, and differentiation in mice that lack or overexpress the cd19 signal transduction molecule, *Immunity*, *3*, 39–50.

Erdmann, T, et al. (2017), Sensitivity to pi3k and akt inhibitors is mediated by divergent molecular mechanisms in subtypes of dlbcl, *Blood*, *130*(3), 310–322.

Erlanson, M, Grönlund, E, Löfvenberg, E, Roos, G, and Lindh, J (1998), Expression of activation markers cd23 and cd69 in b-cell non-hodgkin's lymphoma, *European Journal of Haematology*, *60*, 125–132.

Feist, AM, Herrgard, MJ, Thiele, I, Reed, JL, and Palsson, BO (2009), Reconstruction of biochemical networksyu in microorganisms, *Nat. Rev. Microbiol*, *7*, 129143.

Ferlay, J, Colombet, M, Soerjomataram, I, Mathers, C, Parkin, DM, Pieros, M, Znaor, A, and Bray, F (2019), Estimating the global cancer incidence and mortality in 2018: Globocan sources and methods, *International Journal of Cancer*, *144*(8), 1941–1953.

Ford, CE, Qian, M, Sean, S, Quadir, A, and Ward, RL (2013), The dual role of the novel wnt receptor tyrosine kinase, ror2, in human carcinogenesis, *International Journal of Cancer*, *133*(4), 779–787.

Foulkes, WD, Smith, IE, and Reis-Filho, JS (2010), Triple-negative breast cancer, *New England Journal of Medicine*, *363*(20), 1938–1948.

Fowler, N, and Davis, E (2013), Targeting b-cell receptor signaling: changing the paradigm, *Hematology Am Soc Hematol Educ Program*, *1*, 553560.

Franceschini, A, et al. (2013), String v9.1: protein-protein interaction networks, with increased coverage and integration, *Nucleic Acids Research*, *41*(D1), D808–D815.

Friedman, N (2004), Inferring cellular networks using probabilistic graphical models, *Science*, *303*, 799–805.

Friedman, N, Linial, M, Nachman, I, and Pe'er, D (2000), Using bayesian networks to analyze expression data, *Journal of Comp Biol*, *7*, 601–620.

Fröhlich, H, Fellmann, M, Sueltmann, H, Poustka, A, and Beißbarth, T (2007), Estimating large-scale signaling networks through nested effect models with intervention effects from microarray data, *Bioinformatics*, *8*, 386.

Fruchterman, TMJ, and Reingold, EM (1991), Graph drawing by force-directed placement, *Software: Practice and Experience*, *21*(11), 1129–1164.

Gascoyne, DM, Lyne, L, Spearman, H, Buffa, FM, Soilleux, EJ, and Banham, AH (2017), Vitamin d receptor expression in plasmablastic lymphoma and myeloma cells confers susceptibility to vitamin d, *Endocrinology*, *158*(3), 503–515.

Gauld, SB, Dal Porto, JM, and Cambier, JC (2002), B cell antigen receptor signaling: Roles in cell development and disease, *Science*, *296*(5573), 1641–1642.

Ghanbari, M, Lasserre, J, and Vingron, M (2015), Reconstruction of gene networks using prior knowledge, *BMC Syst Biol*, *9*(84).

Ghosh, R, et al. (2011), Trastuzumab has preferential activity against breast cancers driven by her2 homodimers, *Cancer Research*, *71*(5), 1871–1882.

Gibbs, DL, Gralinski, L, RS, Baric, and McWeeney, SK (2014), Multi-omic network signatures of disease, *Front. Genet*, *4*, 309.

Gilmorec, TD, and Gerondakis, S (2011), The c-rel transcription factor in development and disease, *Genes Cancer*, *2*(7), 695711.

Gold, MR, Scheid, MP, Santos, L, Dang-Lawson, M, Roth, RA, Matsuuchi, L, Duronio, V, and Krebs, DL (1999), The b cell antigen receptor activates the akt (protein kinase b)/glycogen synthase kinase-3 signaling pathway via phosphatidylinositol 3-kinase, *The Journal of Immunology*, *163*(4), 1894–1905.

Goodacre, R, Vaidyanathan, S, Dunn, WB, Harrigan, GG, and Kell, DB (2004), Metabolomics by numbers: acquiring and understanding global metabolite data, *Trends in Biotechnology*, *22*(5), 245–252.

Grumolato, L, et al. (2010), Canonical and noncanonical wnts use a common mechanism to activate completely unrelated coreceptors, *Genes Dev*, *24*(22), 25172530.

Haffty, BG, Yang, Q, Reiss, M, Kearney, T, Higgins, SA, Weidhaas, J, Harris, L, Hait, W, and Toppmeyer, D (2006), Locoregional relapse and distant metastasis in conservatively managed triple negative early-stage breast cancer, *Journal of Clinical Oncology*, *24*(36), 5652–5657.

Hagberg, AA, Schult, DA, and Swart, PJ (2008), Exploring network structure, dynamics, and function using networkx, in *Proceedings of the 7th Python in Science Conference*, p. 1115, Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), Pasadena, CA USA.

Haider, S, and Pal, R (2013), Integrated analysis of transcriptomic and proteomic data, *Curr Genomics*, *2*(14), 91110.

Han, A, Saijo, K, Mecklenbräuker, I, Tarakhovsky, A, and Nussenzweig, M (2003), Bam32 links the b cell receptor to erk and jnk and mediates b cell proliferation but not survival, *Immunity*, *19*, 621632.

Hanahan, D, and Weinberg, RA (2000), The hallmarks of cancer, *Cell*, *1*(100), 5770.

Handcock, MS, Hunter, DR, Butts, CT, Goodreau, SM, and Morris, M (2003), *statnet: Software tools for the Statistical Modeling of Network Data*, Seattle, WA.

Hansen, J andLong L, KD ans Gentry, Gentleman, R, Falcon, S, Hahne, F, and Sarkar, D (2019), Rgraphviz: Provides plotting capabilities for r graph objects, *Leuk Lymphoma*.

Hao, S, Kurosaki, T, and A, August (2003), Differential regulation of nfat and srf by the b cell receptor via a plcγca^{2+}-dependent pathway, *EMBO Journal*, *22*(16), 41664177.

Hardcastle, TJ, and Kelly, KA (2010), bayseq: Empirical bayesian methods for identifying differential expression in sequence count data, *BMC Bioinformatics*, *11*(1), 422.

Hashimoto, A, Okada, H, Jiang, A, Kurosaki, M, Greenberg, S, Clark, EA, and Kurosaki, T (1998), Involvement of guanosine triphosphatases and phospholipase c-γ2 in extracellular signal-regulated kinase, c-jun nh2-terminal kinase, and p38 mitogen-activated protein kinase activation by the b cell antigen receptor, *Journal of Experimental Medicine*, *188*(7), 1287–1295.

Hashimshony, T, Wagner, F, Sher, N, and Yanai, I (2012), Cel-seq: single-cell rna-seq by multiplexed linear amplification, *Cell Reports*, *2*(3), 666–673.

He, J, Sheng, T, Stelter, AA, Li, Ch, Zhang, X, Sinha, M, Luxon, BA, and Xie, J (2006), Suppressing wnt signaling by the hedgehog pathway through sfrp-1, *Journal of Biological Chemistry*, *281*(47), 35,598–35,602.

Henry, C, Quadir, A, Hawkins, NJ, Jary, E, Llamosas, E, Kumar, D, Daniels, B, Ward, RL, and Ford, CE (2015), Expression of the novel wnt receptor ror2 is increased in breast cancer and may regulate both β-catenin dependent and independent wnt signalling, *J Cancer Res Clin Oncol*, *141*(2), 243–254.

Hirosawa, M, Nakahara, M, Otosaka, R, Imoto, A, Okazaki, T, and Takahashi, S (2009), The p38 pathway inhibitor sb202190 activates mek/mapk to stimulate the growth of leukemia cells, *Leukemia Research*, *33*(5), 693–699.

Hochberg, J, El-Mallawany, NK, and Abla, O (2016), Adolescent and young adult non-hodgkin lymphoma, *British Journal of Haematology*, *173*(4), 637–650.

Hommel, G (1988), A stagewise rejective multiple test procedure based on a modified bonferroni test, *Biometrika*, *75*(2), 383–386.

Houser, B (2012), Bio-rad's bio-plex® suspension array system, xmap technology overview, *Archives of Physiology and Biochemistry*, *118*(4), 192–196.

Hu, T, and Li, C (2010), Convergence between wnt-β-catenin and egfr signaling in cancer, *Mol Cancer*, *9*, 236.

Huynh-Thu, VA, Irrthum, A, Wehenkel, L, and Geurts, P (2010), Inferring regulatory networks from expression data using tree-based methods, *PLoS One*, *5*(9), e12,776.

Ideker, TE, Thorsson, V, and Karp, RM (2000), Discovery of regulatory interactions through perturbation: inference and experimental design, *Pac Symp Biocomput*, pp. 305–316.

Irizarry, RA, Bolstad, BM, Collin, F, Cope, LM, Hobbs, B, and Speed, TP (2003), Summaries of affymetrix genechip probe level data, *Nucleic Acids Res*, *31*(4), e15.

Ishikawa, H, Asano, M, Kanda, T, Kumar, S, Gélinas, C, and Ito, Y (1993), Two novel functions associated with the rel oncoproteins: Dna replication and cell-specific transcriptional activation, *Oncogene*, *8*(11), 2889–2896.

Janes, KA, and Lauffenburger, DA (2013), Models of signalling networks - what cell biologists can gain from them and give to them, *Journal of Cell Science*, *126*(9), 1913–1921.

Jansen, R, et al. (2003), A bayesian networks approach for predicting protein-protein interactions from genomic data, *Science*, *302*, 449–453.

Johnson, GL, and Lapadat, R (2002), Mitogen-activated protein kinase pathways mediated by erk, jnk, and p38 protein kinases, *Science*, *298*(5600), 1911–1912.

Jokinen, E, and Koivunen, JP (2015), Mek and pi3k inhibition in solid tumors: rationale and evidence to date, *Therapeutic Advances in Medical Oncology*, *7*(3), 170–180.

Joyce, AR, and Palsson, BO (2006), The model organism as a system: integrating 'omics' data sets, *Nature Reviews Molecular Cell Biology*, *7*, 198210.

Junker, BH, and Schreiber, F (2008), *Analysis of Biological Networks*, Wiley Series in Bioinformatics, John Wiley & Sons, Ltd.

Kallergi, G, Agelaki, S, Kalykaki, A, Stournaras, Ch, Mavroudis, D, and Georgoulias, V (2008), Phosphorylated egfr and pi3k/akt signaling kinases are expressed in circulating tumor cells of breast cancer patients, *Breast Cancer Research*, *10*(5), R80.

Kamburov, A, Pentchev, K, Galicka, H, Wierling, Ch, Lehrach, H, and Herwig, R (2011), Consensuspathdb: toward a more complete picture of cell biology, *Nucleic Acids Research*, *39*, D712–D717.

Kanehisa, M, and Goto, S (2000), Kegg: kyoto encyclopedia of genes and genomes, *Nucleic Acids Res*, *28*, 2730.

Kauffman, SA (1969), Metabolic stability and epigenesis in randomly constructed genetic nets, *Journal of Theoretical Biology*, *22*(3), 437–467.

Kikuchi, A, Yamamoto, H, and Sato, A (2009), Selective activation mechanisms of wnt signaling pathways, *Trends in Cell Biology*, *19*(3), 119–129.

Kitano, H (2002), Systems biology: A brief overview, *Science*, *295*(5560), 1662–1664.

Klemm, F, et al. (2011), β-catenin-independent wnt signaling in basal-like breast cancer and brain metastasis, *Carcinogenesis*, *32*(3), 434–442.

Klinger, B, et al. (2013), Network quantification of egfr signaling unveils potential for targeted combination therapy, *Molecular Systems Biology*, *9*(1).

Klipp, E, and Liebermeister, W (2006), Mathematical modeling of intracellular signaling pathways, *BMC Neurosci*, *7*(1), S10.

Kohno, M, and Pouyssegur, J (2006), Targeting the erk signaling pathway in cancer therapy, *Annals of Medicine*, *38*(3), 200–211.

Komiya, Y, and Habas, R (2008), Wnt signal transduction pathways, *Cell*, *4*(2), 6875.

Koval, A, and Katanaev, VL (2018), Dramatic dysbalancing of the wnt pathway in breast cancers, *Scientific Reportsvolume*, *8*(7329).

Kraus, M, Alimzhanov, MB, Rajewsky, N, and Rajewsky, K (2004), Survival of resting mature b lymphocytes depends on bcr signaling via the igα/β heterodimer, *Cell*, *117*(6), 787–800.

Kurosaki, T, and Kurosaki, M (1997), Transphosphorylation of bruton's tyrosine kinase on tyrosine 551 is critical for b cell antigen receptor function, *J. Biol. Chem*, *272*, 15,595–15,598.

Lam, K-P, Kühn, R, and Rajewsky, K (1997), In vivo ablation of surface immunoglobulin on mature b cells by inducible gene targeting results in rapid cell death, *Cell*, *90*(6), 1073–1083.

Langfelder, P, Zhang, B, and Horvath, S (2007), Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r, *Bioinformatics*, *24*(5), 719–720.

Lee, MJ, Ye, AS, Gardino, AK, Heijink, AM, Sorger, PK, MacBeath, G, and Yaffe, MB (2012), Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signaling networks, *Cell*, *149*(4), 780–794.

Leng, N, et al. (2013), Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments, *Bioinformatics*, *29*(8), 1035–1043.

Lenz, G, and Staudt, LM (2010), Aggressive lymphomas, *N Engl J Med*, p. 362.

Li, B, and Dewey, CN (2011), Rsem: accurate transcript quantification from rna-seq data with or without a reference genome, *BMC Bioinformatics*, *12*, 323.

Li, Y, and Jackson, SA (2015), Gene network reconstruction by integration of prior biological knowledge, *G3: Genes, Genomes, Genetics*, *5*(6), 1075–1079.

Liao, JC, Boscolo, R, Yang, YL, Tran, LM, Sabatti, C, and Roychowdhury, V (2003), Network component analysis: Reconstruction of regulatory signals in biological systems, *Proc Natl Acad Sci*, *100*, 15,52215,527.

Love, MI, Huber, W, and Anders, S (2014), Moderated estimation of fold change and dispersion for rna-seq data with deseq2, *Genome Biology*, *15*(12), 550.

Maathuis, M, Kalisch, M, and Bühlmann, P (2009), Estimating high-dimensional intervention effects from observational data, *Ann Statist*, *37*(6A), 3133–3164.

Malone, JH, and Oliver, B (2011), Microarrays, deep sequencing and the true measure of the transcriptome, *BMC Biology*, *1*(9), 34.

Mannsperger, HA, Gade, S, Henjes, F, Beißbarth, T, and Korf, U (2010), Rppanalyzer: Analysis of reverse-phase protein array data, *Bioinformatics*, *26*(17), 2202–2203.

Markowetz, F, Bloch, J, and Spang, R (2005), Non-transcriptional pathway features reconstructed from secondary effects of rna interference, *Bioinformatics*, *21*, 40264032.

Mathur, Ravi, Schaffer, J, Land, Walker, J Heine, John, M Hernandez, Jonathan, and Yeatman, Timothy (2011), Perturbation and candidate analysis to combat overfitting of gene expression microarray data, *International journal of computational biology and drug design*, *4*, 307–15.

Matys, V, et al. (2006), Transfac and its module transcompel: transcriptional gene regulation in eukaryotes, *Nucleic Acids Res*, *34*, D108D110.

McCarthy, DJ, Chen, Y, and Smyth, GK (2012), Differential expression analysis of multifactor rna-seq experiments with respect to biological variation, *Nucleic Acids Res*, *40*(10), 4288–4297.

McCubrey, JA, et al. (2007), Roles of the raf/mek/erk pathway in cell growth, malignant transformation and drug resistance, *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, *1773*(8), 1263–1284.

McDermott, JE, Wang, J, Mitchell, H, Webb-Robertson, B-J, Hafen, R, Ramey, J, and Rodland, KD (2013), Challenges in biomarker discovery: Combining expert insights with statistical analysis of complex omics data, *Expert Opin Med Diagn*, *7*(1), 3751.

Mendoza, MC, Er, EE, and Blenis, J (2011), The ras-erk and pi3k-mtor pathways: cross-talk and compensation, *Trends in Biochemical Sciences*, *36*(6), 320–328.

Michaut, M, et al. (2016), Integration of genomic, transcriptomic and proteomic data identifies two biologically distinct subtypes of invasive lobular breast cancer, *Scientific Reports*, *6*, 18,517.

Musgrove, EA (2004), Wnt signalling via the epidermal growth factor receptor: a role in breast cancer?, *Breast Cancer Res*, *6*(2), 65–68.

Ng, SB, et al. (2009), Targeted capture and massively parallel sequencing of twelve human exomes, *Nature*, *7261*(461), 272276.

Nishimura, D (2001), Biocarta, *Biotechnol. Softw. Internet Rep*, *2*, 117120.

Nishizuka, S, et al. (2003), Proteomic profiling of the nci-60 cancer cell lines using new high-density reverse-phase lysate microarrays, *Proc Natl Acad Sci U S A*, *100*(24), 14,22914,234.

Nolz, JC, Tschumper, RC, Pittner, BT, Darce, JR, Kay, NE, and Jelinek, DF (2005), Zap-70 is expressed by a subset of normal human b-lymphocytes displaying an activated phenotype, *Leukemia*, *19*, 10181024.

O'Brien, KM, Cole, SR, Tse, CK, Perou, CM, Carey, LA, Foulkes, WD, Dressler, LG, Geradts, J, and Millikan, RC (2010), Intrinsic breast tumor subtypes, race, and long-term survival in the carolina breast cancer study, *Clin Cancer Res*, *16*(24), 61006110.

Oishi, I, et al. (2003), The receptor tyrosine kinase ror2 is involved in non-canonical wnt5a/jnk signalling pathway, *Genes to Cells*, *8*(7), 645–654.

Orian-Rousseau, V, and Schmitt, M (2015), Cd44 regulates wnt signaling at the level of lrp6, *Molecular & Cellular Oncology*, *2*(3), e995,046.

O'Toole, SA, et al. (2011), Hedgehog overexpression is associated with stromal interactions and predicts for poor outcome in breast cancer, *Cancer Research*, *71*(11), 4002–4014.

Park, H, Wahl, MI, Afar, DE, Turck, CW, Rawlings, DJ, Tam, C, Scharenberg, AM, Kinet, JP, and Witte, ON (1996), Regulation of btk function by a major autophosphorylation site within the sh3 domain, *Immunity*, *3*(3), 515–525.

Park, JH, Lee, SY, Kim, TY, and Kim, HU (2008), Application of systems biology for bioprocess development, *Trends Biotechnol*, *26*, 404412.

Paweletz, CP, et al. (2001), Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front, *Oncogene*, *20*, 19811989.

Pearl, J (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers Inc, San Francisco, CA, USA.

Peck, AR, et al. (2011), Loss of nuclear localized and tyrosine phosphorylated stat5 in breast cancer predicts poor clinical outcome and increased risk of antiestrogen therapy failure, *Journal of Clinical Oncology*, *29*(18), 2448–2458.

Perou, CM, et al. (2000), Molecular portraits of human breast tumors, *Nature*, *406*, 747752.

Picotti, P1, and Aebersold, R (2012), Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions, *Nat Methods*, *6*(9), 555–566.

Pirkl, M, Hand, E, Kube, D, and Spang, R (2016), Analyzing synergistic and non-synergistic interactions in signalling pathways using boolean nested effect models, *Bioinformatics*, *32*(6), 893–900.

Pukrop, T., Klemm, F., Hagemann, Th., Gradl, D., Schulz, M., Siemes, S., Trümper, L., and Binder, C. (2006), Wnt 5a signaling is critical for macrophage-induced invasion of breast cancer cell lines, *Proceedings of the National Academy of Sciences*, *103*(14), 5454–5459.

Rapaport, F, Khanin, R, Liang, Y, Pirun, M, Krek, A, Zumbo, P, Mason, CE, Socci, ND, and Betel, D (2013), Comprehensive evaluation of differential gene expression analysis methods for rna-seq data, *Genome Biology*, *14*(9), 3158.

Reshetova, P, Smilde, AK, van Kampen, AHC, and Westerhuis, JA (2014), Use of prior knowledge for the analysis of high-throughput transcriptomics and metabolomics data, *BMC Systems Biology*, *8*(2), S2.

Reth, M, and Brummer, T (2004), Feedback regulation of lymphocyte signalling, *Nature Reviews Immunology*, *4*(4), 269.

Reth, M, and Wienands, J (1997), Initiation and processing of signals from the b cell antigen receptor, *Nucleic Acids Research*, *15*, 453–479.

Rice, JJ, Tu, Y, and Stolovitzky, G (2005), Reconstructing biological networks using conditional correlation analysis, *Bioinformatics*, *21*(6), 765–773.

Ritchie, ME, Phipson, B, Wu, D, Hu, Y, Law, CW, Shi, W, and Smyth, GK (2015), limma powers differential expression analyses for rna-sequencing and microarray studies, *Nucleic Acids Research*, *7*(43), e47.

Roarty, K, Pfefferle, AD, Creighton, CJ, Perou, CM, and Rosen, JM (2017), Ror2-mediated alternative wnt signaling regulates cell fate and adhesion during mammary tumor progression, *Oncogene*, *36*(43), 59585968.

Robinson, MD, and Smyth, GK (2008), Small-sample estimation of negative binomial dispersion, with applications to sage data, *Biostatistics,*, *9*(2), 321–332.

Robinson, MD, McCarthy, DJ, and Smyth, GK (2010), edger: a bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics*, *26*(1), 139–140.

Rosso, SB, and Inestrosa, NC (2013), Wnt signaling in neuronal maturation and synaptogenesis, *Front Cell Neurosci*, *7*, 103.

Rouprêt, M, et al. (2008), A comparison of the performance of microsatellite and methylation urine analysis for predicting the recurrence of urothelial cell carcinoma, and definition of a set of markers by bayesian network analysis, *BJU International*, *101*, 1448–1453.

Sachs, K, Perez, O, Pe'er, D, Lauffenburger, DA, and Nolan, GP (2005), Causal protein-signaling networks derived from multi-parameter single-cell data, *Science*, *308*, 523–529.

Sadeghi, A, and Fröhlich, H (2013), Steiner tree methods for optimal sub-network identification: An empirical study, *Bioinformatics*, *37*, D674D679.

Saez-Rodriguez, Julio, Alexopoulos, Leonidas G., and Stolovitzky, Gustavo (2011), Setting the standards for signal transduction research, *Science Signaling*, *4*(160), pe10.

Saito, K, Tolias, KF, Saci, A, Koon, HB, Humphries, LA, Scharenberg, A, Rawlings, DJ, Kinet, JP, and Carpenter, CL (2003), Btk regulates ptdins-4,5-p2 synthesis: Importance for calcium signaling and pi3k activity, *Immunity*, *19*(5), 669–677.

Sansone, P, and Bromberg, J (2012), Targeting the interleukin-6/jak/stat pathway in human malignancies, *Journal of Clinical Oncology*, *30*(9), 1005–1014.

Sato, A, Yamamoto, H, Sakane, H, Koyama, H, and Kikuchi, A (2010), Wnt5a regulates distinct signalling pathways by binding to frizzled2, *EMBO J*, *1*(29), 4154.

Schaefer, CF, Anthony, K, Krupa, S, Buchoff, J, Day, M, Hannay, T, and Kenneth, HB (2009), Pid: the pathway interaction database, *Nucleic Acids Res*, *144*(14), D674D679.

Schena, M, Shalon, D, Davis, RW, and Brown, PO (1995), Quantitative monitoring of gene expression patterns with a complementary dna microarray, *Science*, *270*(5235), 467–470.

Schlange, T, Matsuda, Y, Lienhard, S, Huber, A, and Hynes, Nancy E (2007), Autocrine wnt signaling contributes to breast cancer cell proliferation via the canonical wnt pathway and egfr transactivation, *Breast Cancer Research*, *9*(R63).

Schurch, NJ, et al. (2016), How many biological replicates are needed in an rna-seq experiment and which differential expression tool should you use?, *RNA (New York, N.Y.)*, *22*(6), 839–851.

Schwanhäusser, B, Busse, D, Li, N, Dittmar, G, Schuchhardt, J, Wolf, J, Chen, W, and Selbach, M (2011), Global quantification of mammalian gene expression control, *Nature*, *473*(7347), 337–342.

Schwartz, GW, Petrovic, J, Zhou, Y, and Faryabi, RB (2018), Differential integration of transcriptome and proteome identifies pan-cancer prognostic biomarkers, *Front Genet*, *9*, 209.

Schwarz, G (1978), Estimating the dimension of a model, *The Annals of Statistics*, *6*, 461–464.

Segal, E, Shapira, M, Regev, A, Pe'er, D, Botstein, D, Koller, D, and Friedman, N (2003), Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data, *Nature Genetics*, *34*, 166176.

Sever, R, and Brugge, JS (2015), Signal transduction in cancer, *Cold Spring Harb Perspect Med*, *4*(5).

Shannon, P, Markiel, A, Ozier, O, Baliga, NS, Wang, JT, Ramage, D, Amin, N, Schwikowski, B, and Ideker, T (2003), Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Research*, *13*(11), 2498–2504.

Sharma, M, Chuang, WW, and Sun, Z (2002), Phosphatidylinositol 3-kinase/akt stimulates androgen pathway through gsk3β inhibition and nuclear β-catenin accumulation, *Journal of Biological Chemistry*, *277*(34), 30,935–30,941.

Shaw, RJ, and Cantley, LC (2006), Ras, pi(3)k and mtor signalling controls tumour cell growth, *Nature*, *441*, 424430.

Shen, R, Olshen, AB, and Ladanyi, M (2009), Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis, *Bioinformatics*, *25*, 29062912.

Shin, S-Y, Rath, O, Choo, S-M, Fee, F, McFerran, B, Kolch, W, and Cho, K-H (2009), Positive- and negative-feedback regulations coordinate the dynamic behavior of the ras-raf-mek-erk signal transduction pathway, *Journal of Cell Science*, *122*(3), 425–435.

Simons, M, and Mlodzik, M (2008), Planar cell polarity signaling: From fly development to human disease, *Annual Review of Genetics*, *42*(1), 517–540.

Sitte, M, Menck, K, Wachter, A, Reinz, E, Korf, U, Wiemann, S, Bleckmann, A, and Beißbarth, T (2019), Reconstruction of different modes of wnt dependent protein networks from time series protein quantification, *Stud Health Technol Inform*, *267*, 175–180.

Smyth, GK (2004), Linear models and empirical bayes methods for assessing differential expression in microarray experiments, *Statistical Applications in Genetics and Molecular Biology*, *3*(1), 1–25.

Smyth, GK (2005), limma: Linear models for microarray data, in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Statistics for Biology and Health*, edited by Gentleman, R, Carey, VJ, Huber, W, Irizarry, RA, and Dudoit, S, pp. 397–420, Springer New York.

Soneson, C, and Delorenzi, M (2013), A comparison of methods for differential expression analysis of rna-seq data, *BMC Bioinformatics*, *14*(1), 91.

Sørlie, Therese (2004), Molecular portraits of breast cancer: tumour subtypes as distinct disease entities, *European Journal of Cancer*, *40*(18), 2667–2675.

Sørlie, T, et al. (2001), Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications, *Proceedings of the National Academy of Sciences*, *98*(19), 10,869–10,874.

Spurrier, B, Ramalingam, S, and Nishizuka, S (2008), Reverse-phase protein microarrays for cell signaling analysis, *Nature Protocols*, *3*, 59–71.

Stark, C, Breitkreutz, B-J, Reguly, T, Boucher, L, Breitkreutz, A, and Tyers, M (2006), Biogrid: a general repository for interaction datasets, *Nucleic Acids Research*, *34*, D535–D539.

Stavrakas, V, Melas, I N, Sakellaropoulos, T, and Alexopoulos, LG (2015), Network reconstruction based on proteomic data and prior knowledge of protein connectivity using graph theory, *PLoS ONE*, *10*(5), 1–17.

Steelman, LS, et al. (2011), Roles of the raf/mek/erk and pi3k/pten/akt/mtor pathways in controlling growth and sensitivity to therapy-implications for cancer and aging, *Aging*, *3*(3), 192–222.

Steen, H, Jebanathirajah, JA, Rush, J, Morrice, N, Kirschner, and Marc, W (2006), Phosphorylation analysis by mass spectrometry, *Molecular & Cellular Proteomics*, *5*(1), 172–181.

Sugiyama, K, Tagawa, S, and Toda, M (1981), Methods for visual understanding of hierarchical system structures, *IEEE Transactions onSystems, Man, and Cybernetics*, *2*(11), 109–125.

Swerdlow, SH, Campo, E, Harris, NL, Jaffe, ES, Pileri, SA, Stein, H, Thiele, J, and Vardiman, JW (2008), *World Health Organization Classification of Tumours of Haematopoietic and Lymphoid Tissues*, vol. 2, IARC Pres, Lyon, France.

Taylor-Fishwick, DA, and Siegel, JN (1995), Raf-1 provides a dominant but not exclusive signal for the induction of cd69 expression on t cells, *Eur. J. Immunol*, *25*, 3215–3221.

Teiten, MH, Blasius, R, Morceau, F, Diederich, M, and Dicato, M (2007), Drug discovery technologies, in *Comprehensive Medicinal Chemistry II*, edited by JB, Taylor, pp. 189–214, Elsevier Science.

Teixeira, Al, Ribeiro, R, Morais, A, Lobo, F, Fraga, A, Pina, F, Calais-da Silva, FM, Calais-da Silva, FE, and Medeiros, R (2009), Combined analysis of egf+61g¿a and tgfb1+869t¿c functional polymorphisms in the time to androgen independence and prostate cancer susceptibility, *The Pharmacogenomics Journal*, *9*, 341–346.

Terfve, C, and Saez-Rodriguez, J (2012), Modeling signaling networks using high-throughput phospho-proteomics, in *Advances in Systems Biology*, edited by Goryanin, II and Goryachev, AB, pp. 19–57, Springer New York, New York, NY.

Therneau, TM (2015), *A Package for Survival Analysis in S*, version 2.38.

Tran, LM, Brynildsen, MP, Kao, KC, Suen, JK, and Liao, JC (2005), gnca: A framework for determining transcription factor activity based on transcriptome: Identifiability and numerical implementation, *Metabolic Engineering*, *7*, 128141.

Tresch, A, and Markowetz, F (2008), Structure learning in nested effects models, *Statistical Applications in Genetics and Molecular Biology*, *12*(7), 14,857–70.

Tukey, JW (1977), Some thoughts on clinical trials, especially problems of multiplicity, *Science*, *198*(4318), 679–684.

Tuncbag, N, Braunstein, A, A nd Pagnani, Huang, SS, Chayes, J, Borgs, C, Zecchina, R, and Fraenkel, E (2013), Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem, *J Comput Biol*, *20*(2), 124136.

Ummanni, R, Mannsperger, HA, Sonntag, J, Oswald, M, Sharma, AK, Knig, R, and Korf, U (2014), Evaluation of reverse phase protein array (rppa)-based pathway-activation profiling in 84 non-small cell lung cancer (nsclc) cell lines as platform for cancer proteomics and biomarker discovery, *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1844*(5), 950–959.

van Amerongen, R (2012), Alternative wnt pathways and receptors, *Cold Spring Harb Perspect Biol*, *4*(10), a007,914.

van Someren, EP, Vaes, BLT, Steegenga, WT, Sijbers, AM, Dechering, KJ, and Reinders, MJT (2006), Least absolute regression network analysis of the murine osteoblast differentiation network, *Bioinformatics*, *22*(4), 477–484.

Veiga, DF, Vicente, FF, Grivet, M, de la Fuente, A, and Vasconcelos, AT (2007), Genome-wide partial correlation analysis of esche-richia coli microarray data, *Genet Mol Res*, *20*(18), 730742.

Verkaar, F, and Zaman, GJ (2010), A model for signaling specificity of wnt/frizzled combinations through co-receptor recruitment, *FEBS Lett*, *18*(584), 3850–4385.

Vivanco, I, and Sawyers, CL (2002), The phosphatidylinositol 3-kinaseakt pathway in human cancer, *Nature Reviews Cancer*, *2*, 489501.

von der Heyde, S, Sonntag, J, Kaschek, D, Bender, C, Bues, J, Wachter, A, Timmer, J, Korf, U, and Beißbarth, T (2014), Rppanalyzer toolbox: an improved r package for analysis of reverse phase protein array data, *Biotechniques*, *57*(3), 125–135.

von der Heyde, S, Sonntag, J, Kramer, F, Bender, Ch, Korf, U, and Beißbarth, T (2016), Reconstruction of protein networks using reverse-phase protein array data, in *Statistical Analysis in Proteomics*, edited by Jung, K, pp. 227–246, Springer New York, New York, NY.

Wachter, A, and Beißbarth, T (2015), pwomics: an r package for pathway-based integration of time-series omics data using public database knowledge, *Bioinformatics*, *15*, 3072–3074.

Wang, K, Grivennikov, SI, and Karin, M (2013), Implications of anti-cytokine therapy in colorectal cancer and autoimmune diseases, *Annals of the Rheumatic Diseases*, *72*, 100–103.

Wang, W-S, Chen, P-M, Chiou, T-J, Liu, J-H, Lin, J-K, Lin, T-C, Wang, H-S, and Su, Y (2007), Epidermal growth factor receptor r497k polymorphism is a favorable prognostic factor for patients with colorectal carcinoma, *Clinical Cancer Research*, *13*(12), 3597–3604.

Werhli, AV, and Husmeier, D (2008), Gene regulatory network reconstruction by bayesian integration of prior knowledge and/or different experimental conditions, *J Bioinform Comput Biol*, *6*(3), 543572.

Werner, M, Hobeika, E, and Jumaa, H (2010), Role of pi3k in the generation and survival of b cells, *Immunological Reviews*, *237*(1), 55–71.

Westerhoff, HV, and Palsson, BO (2004), The evolution of molecular biology into systems biology, *Nat Biotechnol*, *10*(22), 1249–1252.

Wilson, WH (2013), Treatment strategies for aggressive lymphomas: what works?, *Hematology Am Soc Hematol Educ Program*, pp. 584–590.

Yamanashi, Y, et al. (1997), Role of tyrosine phosphorylation of hs1 in b cell antigen receptor-mediated apoptosis, *Journal of Experimental Medicine*, *185*, 1387–1392.

Yang, K, et al. (2011), Fzd7 has a critical role in cell proliferation in triple negative breast cancer, *Oncogene*, *20*, 44374446.

Yang, L, et al. (2016), The evolving roles of canonical wnt signaling in stem cells and tumorigenesis: implications in targeted cancer therapies, *Laboratory Investigation*, *95*, 116136.

Yang, YH, Dudoit, S, Luu, P, Lin, DM, Peng, V, Ngai, J, and Speed, TP (2002), Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Res*, *30*(4), e15.

Young, D, Stark, J, and Kirschner, D (2008), Systems biology of persistent infection: tuberculosis as a case study, *Nat. Rev. Microbiol*, *6*, 520528.

Yu, J, Smith, VA, Wang, PP, Hartemink, AJ, and Jarvis, ED (2004), Advances to bayesian network inference for generating causal networks from observational biological data, *Bioinformatics*, *20*, 3594–3603.

Yuan, JS, Galbraith, DW, Dai, SY, Griffin, P, and Stewart, CN Jr (2008), Plant systems biology comes of age, *Trends Plant Sci*, *13*, 165171.

Yuen, HF, et al. (2012), Impact of oncogenic driver mutations on feedback between the pi3k and mek pathways in cancer cells, *Trends Plant Sci*, *4*(32), 413–422.

Zak, DE, and Aderem, A (2009), Systems biology of innate immunity, *Immunol. Rev*, *1*, 264282.

Zeller, C, Fröhlich, H, and Tresch, A (2009), A bayesian network view on nested effects models, *EURASIP Journal on Bioinformatics and Systems Biology*, *227*.

Zeng, CM, Chen, Z, and Fu, L (2018), Frizzled receptors as potential therapeutic targets in human cancers, *Int J Mol Sci*, *19*(5), 1543.

Zhang, B, et al. (2014), Proteogenomic characterization of human colon and rectal cancer, *Nature*, *513*, 382387.

Zhang, L, Wei, Q, Mao, L, Liu, W, Mills, GB, and Coombes, K (2009), Serial dilution curve: a new method for analysis of reverse phase protein array data, *Bioinformatics*, *25*(5), 650654.

Zhimin, L, Sourav, G, Zhiyong, W, and Tony, H (2003), Downregulation of caveolin-1 function by egf leads to the loss of e-cadherin, increased transcriptional activity of β-catenin, and enhanced tumor cell invasion, *Cancer Cell*, *4*(6), 499–515.

Zhu, J, Zhang, B, and Schadt, EE (2008), A systems biology approach to drug discovery, *Adv. Genet*, *60*.

Zuo, Y, Yu, G, Tadesse, MG, and Ressom, HW (2014), Biological network inference using low order partial correlation, *Methods*, *69*(3), 266–273, recent development in bioinformatics for utilizing omics data.

Zyprych-Walczak, J, Szabelska, A, Handschuh, L, Górczak, K, Klamecka, K, Figlerowicz, M, and I, Siatkowski (2015), The impact of normalization methods on rna-seq data analysis, *Biomed Res Int*, *621690*.

DECLARATION

Hereby, I declare that I have composed the presented thesis entitled “Network Based Integration of Proteomic and Transcriptomic Data: Study of BCR and WNT11 Signaling Pathways in Cancer Cells“ independently on my own and without any other resources than the ones indicated.

____________________ ____________________

Date Maren Sitte