Pathway and network analyses in context of Wnt signaling in breast cancer

Dissertation
for the award of the degree

Doctor of Philosophy

Division of Mathematics and Natural Sciences
of the Georg-August-Universität Göttingen
within the doctoral program Molecular Biology of Cells
of the Georg-August-University School of Science (GAUSS)

submitted by
Michaela Bayerlová
born in
Košice, Slovakia

Göttingen, 2015

Thesis Committee:

Prof. Dr. Tim Beißbarth
Department of Medical Statistics
University Medical Center Göttingen
Prof. Dr. Burkhard Morgenstern
Department of Bioinformatics
Institute of Microbiology and Genetics
University of Göttingen
Prof. Dr. med. Tobias Pukrop
Department of Hematology and Medical Oncology
University Medical Center Göttingen

Members of the Examination Board:
Referee: Prof. Dr. Tim Beißbarth
2nd Referee: Prof. Dr. Burkhard Morgenstern

Further members of the Examination Board:
Prof. Dr. Edgar Wingender
Prof. Dr. Gregor Bucher
Prof. Dr. Steven Johnsen
Prof. Dr. med. Heidi Hahn

Date of oral examination: 14th of January 2016

Abstract

A complex network of interplaying signaling pathways governs cell behavior and phenotype. Wnt signaling pathways are part of this network and play an important role in embryonic development as well as in carcinogenesis. In particular, non-canonical Wnt signaling is considered critical for breast cancer cell proliferation and migration. However, specific outcomes of distinct Wnt signaling pathways are still poorly understood.

To better characterize these processes, gene expression responses of aberrant Wnt signaling can be quantified by expression profiling, and further analyzed using various bioinformatic approaches. In particular, pathway enrichment and network integration are effective strategies to obtain a comprehensive interpretation of the results of differential expression analysis.

Enrichment analysis is a widely used tool to detect pathways significantly altered between two experimental conditions. Before applying this approach in the Wnt signaling context, enrichment methods were evaluated in an extensive comparative study to assess the contribution of pathway structure integration into the enrichment analysis. Standard gene-set methods were compared against pathway topology-based methods in multiple simulation scenarios and on benchmark data. These results as well as a critical consideration of methodological principles suggest that simple gene-set enrichment methods are favorable.

In order to elucidate the role of Wnt signaling in aggressive breast cancer, changes in the expression profiles of breast cancer cells after over-expression of the non-canonical Wnt receptor Ror2 were analyzed. Over-expression resulted in increased cell invasion and over 2000 differential target genes of this perturbation were identified. These targets were further placed into the context of known signaling pathways and molecular networks.

To this end, the public Wnt pathway knowledge was assembled into signaling network models representing distinct Wnt pathways. Subsequently, the Wnt networks were analyzed with regard to their structural properties, and also utilized for the analysis of targets. Results of the enrichment analysis suggest that the Ror2 over-expression activates non-canonical Wnt signaling, whereas canonical Wnt signaling appears not to be affected. Furthermore, integration of targets with the non-canonical Wnt network revealed a differentially regulated module of the non-canonical Wnt signaling and its topologically essential elements were identified. Moreover, target hubs were determined by integration with protein-protein interaction network.

To validate whether the identified Wnt module genes and hub genes are indeed associated with the observed phenotype of increased cell invasion, the results were translated into a clinical context of metastatic breast cancer patients. These two gene lists were utilized as signatures to test prognosis of

metastasis-free survival. Both signatures as well as multiple individual genes were shown to be significantly associated with breast cancer outcome; including several genes that have been previously reported to be potential therapeutic targets or biomarkers.

In conclusion, gene set enrichment analysis as well as bioinformatic approaches derived from network theory were demonstrated to be powerful tools for analyzing the complex gene expression patterns of breast cancer cells. These strategies were shown to provide valuable insights into signaling processes underlying breast cancer phenotypes, particularly highlighting the importance of the non-canonical Wnt pathway in aggressive breast cancer.

Acknowledgements

I would like to express my deep thanks to my scientific advisors Prof. Dr. Tim Beißbarth and Dr. med. Annalen Bleckmann for supervision, support, guidance throughout the projects, and sharing their expertise and their time.

I am very grateful to my colleagues, current as well as former ones: Silvia, Frank, Andreas, Manuel, Klaus, Stephan, Alex, Jochen, Astrid, Xenia, and Julia for fruitful (not always) scientific discussions, inspiring ideas, sunny lunch breaks, and for creating such a great office spirit. Special thanks go to Annalen, Silvia, Astrid, and Julia for the ladies nights and for reading things. I am also very grateful to Lucy for proofreading this thesis.

I wish to acknowledge my collaborators Dr. med. Florian Klemm and Prof. Dr. med. Tobias Pukrop for the contributions to the projects and especially for sharing their great research enthusiasm. Further, I would like to acknowledge the thesis committee members Prof. Dr. Tim Beißbarth, Prof. Dr. Burkhard Morgenstern, and Prof. Dr. med. Tobias Pukrop for constructive discussions.

A big thank goes to Janka, Dorota, Asia, and Björn for their love.

Another big thank goes to my parents also for their love.

Contents

List of Figures **x**

List of Tables **xi**

Abbreviations **xiii**

1 Introduction **1**
1.1 Wnt signaling pathways 2
1.2 Breast cancer 3
1.2.1 Wnt signaling in breast cancer 4
1.3 Gene expression analysis 4
1.4 Representation of pathway knowledge 6
1.5 Enrichment analysis 7
1.5.1 Gene-set enrichment approach 7
1.5.2 Pathway topology-based approach 9
1.6 Molecular networks 10
1.6.1 Types of networks 10
1.6.2 Network analysis 11
1.6.3 Network integration 13
1.7 Aims and organization of the thesis 14

2 Materials and Methods **19**
2.1 Materials 19
2.1.1 Public microarray data 19
2.1.1.1 Patient datasets 19
2.1.1.2 Benchmark datasets 20
2.1.2 Newly generated RNA-Seq data 20
2.1.3 Pathway and network databases 22
2.2 Methods 24

2.2.1 Gene-set enrichment methods . . . 24
2.2.2 Pathway topology-based methods . . . 25
2.2.3 Parsing pathway knowledge from databases . . . 28
2.2.4 Differential analysis . . . 30
2.2.5 Network-based analyses . . . 31
2.2.6 Clustering and survival analysis . . . 32
2.3 Simulations . . . 33
2.3.1 Studies with different pathway input . . . 33
2.3.2 Variable parameters . . . 34
2.3.3 Topology designs for pathway deregulation . . . 36
2.3.4 Single simulation run . . . 36
2.3.5 Evaluation . . . 37

3 Results **39**
3.1 Parsed pathways . . . 39
3.2 Comparison of enrichment methods . . . 40
3.2.1 Simulation studies . . . 41
3.2.1.1 Study 1: with original pathways . . . 41
3.2.1.2 Study 2: with non-overlapping pathways . . . 42
3.2.1.3 Overall performance in simulations . . . 45
3.2.2 Performance in benchmark data . . . 45
3.3 Wnt networks . . . 46
3.3.1 Network properties . . . 47
3.3.2 Clustered Wnt subnetwork . . . 48
3.4 Sequenced cell line and perturbation targets . . . 50
3.4.1 Enrichment of Ror2 targets in Wnt gene sets . . . 53
3.5 Integration of targets with networks . . . 54
3.5.1 Non-canonical Wnt module . . . 54
3.5.2 PPI network and hubs . . . 55
3.6 Breast cancer metastasis-free survival study . . . 58
3.6.1 Patient cohort and subtype prediction . . . 58
3.6.2 Wnt module genes as prognostic signature . . . 59
3.6.3 PPI hub genes as prognostic signature . . . 64

4 Discussion **67**
4.1 Parsing and representing pathway knowledge . . . 67
4.2 Enrichment methods . . . 68

4.3 Wnt networks 73
4.4 Targets and network integration 76
4.5 Breast cancer metastasis and prognostic genes 79

5 Conclusions and Outlook **85**

Appendix **89**

References **101**

List of Figures

1.1 Concept of data integration 18

2.1 Parsing pathway data 30
2.2 Topology designs 37

3.1 Distribution of size of KEGG pathways 40
3.2 Simulation study 1 43
3.3 Simulation study 2 44
3.4 Overall performance 45
3.5 Performance in benchmark data 46
3.6 Wnt networks 48
3.7 Subnetwork of reviewed Wnt genes 49
3.8 Invasion assay of MCF-7 cells 50
3.9 Targets of Wnt5a stimulation 52
3.10 Common Ror2 targets 52
3.11 Non-canonical Wnt module 55
3.12 Hub targets in PPI network 56
3.13 Predicted breast cancer subtypes for patients 59
3.14 Wnt module signature: clustering 60
3.15 Wnt module signature: KM curves 61
3.16 Wnt module signature in subtypes 62
3.17 Wnt module in MFS context 63
3.18 PPI hubs as gene signature 65

List of Tables

2.1 Benchmark datasets 23
2.2 Pathway databases 23
2.3 Enrichment methods 24
2.4 Simulation settings 34
2.5 Measures for performance evaluation 38

3.1 RNA-Seq of MCF-7 cell line 51
3.2 Differential analysis of RNA-Seq 52
3.3 ORA of Ror2 targets in Wnt gene sets 53
3.4 List of Hubs 57
3.5 Patients with breast cancer metastasis 58

1 Appendix Table 1 90
2 Appendix Table 2 91
3 Appendix Table 3 92
4 Appendix Table 4 95
5 Appendix Table 5 100

Abbreviations

basal	-	basal-like
BioPAX	-	Biological Pathways Exchange
ctl	-	control
DC	-	detection call
DEGs	-	differentially expressed genes
Dvl	-	Dishevelled
ER	-	estrogen receptor
FCS	-	functional class scoring
FDR	-	false discovery rate
FE	-	Fisher's exact
Fz	-	Frizzled
GEO	-	Gene Expression Omnibus
GS	-	gene-set
HER2	-	epidermal growth factor receptor
Her2	-	HER2-enriched
HGNC	-	HUGO Gene Nomenclature Committee
HR	-	hazard ration
KM	-	Kaplan-Meier
KS	-	Kolmogorov-Smirnov
logFC	-	logarithm of fold-change
lumA	-	luminal A
lumB	-	luminal B
MFS	-	metastasis-free survival
mRNA	-	messenger RNA
normal	-	normal-breast-like

Abbreviations

ORA	-	over-representation analysis 41
PCP	-	Planar cell polarity 49
PPI	-	protein-protein interaction 87
PR	-	progesterone receptor 3
PT-based	-	pathway topology-based 85
RNA-Seq	-	deep sequencing of RNA 5
WRS	-	Wilcoxon rank sum 41

Introduction

Basic functions of cells are governed by a complex network of interconnected signaling pathways, representing a cellular communication system. A signal is mediated via a ligand binding cell-surface receptor, which subsequently triggers cascades of interactions and biochemical reactions towards transcription factors. The transcription factors in the nucleus then regulate expression of their target genes. The target messenger RNA (mRNA) is further processed and translated, resulting in a finely controlled synthesis of new proteins. Profiles of expressed genes and proteins in the cells determine their phenotype and responses to the environment, such as cell division, differentiation, and apoptosis.

Thus, the behavior of cells is coordinated by signaling networks that translate external signals via gene expression changes into appropriate responses. In case of aberrant signal transduction, the inappropriate pathway regulation can result in diseases, such as cancer or immune disorders.

Within this thesis I focus on the module of this cellular network, which represents Wnt signaling pathways. Deregulation of these pathways has been implicated in a number of cancers including breast tumors. I aim to create network models of distinct Wnt pathways and to elucidate their roles in different breast cancer types using various bioinformatic approaches. Furthermore, I am interested in evaluating a set of bioinformatic methods that are based on different pathway representation strategies. To that end, I first provide the reader with the general background on Wnt signaling and breast cancer, as well as with a comprehensive overview of bioinformatic approaches for gene expression, pathway and network analyses. In the 1.7 *Aims and organization of the thesis* section of the *Introduction* chapter I summarize the aims of this thesis in more detail and describe the organization structure of the thesis.

1.1 Wnt signaling pathways

Wnt proteins are secreted ligands activating a complex mechanism of signal transduction via multiple pathways: the canonical β-catenin-dependent pathway and several non-canonical β-catenin-independent pathways. These Wnt cascades play an important role in embryonic development processes as well as in carcinogenesis (Kahn, 2014).

Activity of the canonical Wnt pathway is defined by the translocation of β-catenin into the nucleus, where it acts as a co-activator of transcription. In the *off-state* no Wnt ligand is present and β-catenin is targeted and degraded by a destruction complex which includes Axin-1, APC, and GSK3B, among others. In the *on-state* a Wnt ligand binds to a Frizzled (Fz) receptor and its co-receptors LRP5/6 activating intracellular Dishevelled (Dvl). This in turn leads to inactivation of the β-catenin destruction complex, which cause that β-catenin is stabilized and relocated into the nucleus. There it functions as a co-activator of TCF/LEF transcription factors, triggering the expression of specific target genes (Yost et al., 1996; Miller et al., 1999). This transcriptional activity determines cell fate decisions, survival, and proliferation.

Moreover, several alternative non-canonical Wnt signaling routes exist, which operate independently from the β-catenin-mediated transcription. These β-catenin-independent cascades can be separated into the Wnt/Planar cell polarity (PCP), Wnt/Ca^{2+} and Wnt/Ror signaling branches. However, these three signaling branches exhibit a considerable degree of overlap (De, 2011). Non-canonical Wnt proteins can also bind to Fz receptors and downstream Dvl is then able to transduce the signal into multiple routes: Via small GTPases, such as RhoA, Rac and CDC42, the cell polarity is regulated. Alternatively, calcium flux is induced activating CaMKII and JNK kinases resulting in transcriptional activity of NFAT and AP-1. Moreover, also non-frizzled, tyrosine kinase-like orphan receptors Ror1 and Ror2 can mediate Wnt signals. Coupling of Wnt5a to Ror2 activates the Ca^{2+}/JNK pathway as well as traffics the signal towards PCP. The outcome of non-canonical Wnt signaling in general is associated with differentiation, changes in cell motility, cytoskeletal rearrangement, and invasiveness (Oishi et al., 2003; Moon et al., 2004; De, 2011; Anastas and Moon, 2013).

In summary, Wnt signaling pathways and sub-pathways should not be seen as straightforward linear cascades, but rather be considered as complex overlapping signaling networks, in which combinatorial activation of components results in context-specific expression responses (Amerongen and Nusse, 2009).

1.2 Breast cancer

Cancer often arises due to defects in important signaling pathways. The changes in gene expression and protein function subsequently alter the phenotype of the cells. Breast cancer is a heterogeneous disease with diverse outcomes in terms of progression, recurrence, metastases formation and overall survival.

Based on the status of hormone estrogen receptor (ER), progesterone receptor (PR) and epidermal growth factor receptor (HER2), three clinically prognostic groups have been established: hormone receptor (ER or PR) positive, HER2 positive and triple negative breast tumors. Whereas the ER or PR positive group has better prognosis and the lower relapse rate, the triple negative group has very unfavorable prognosis with increased risk of local relapse as well as of development of distant metastases (Sørlie et al., 2001; Rakha et al., 2007; Onitilo et al., 2009).

With advent of large-scale gene expression profiling technologies multiple gene signatures have been established for further breast cancer classification. Using 534 *intrinsic* genes and latter suggested *pam50* gene signature, five molecular subtypes associated with different prognosis have been defined: basal-like (basal), HER2-enriched (Her2), luminal A (lumA), luminal B (lumB), and normal-breast-like (normal) subtype. Relating them to the clinical groups, lumA and lumB subtypes correspond to ER positive tumors, whereas the basal subtype correlates with triple negative breast cancers (Perou et al., 2000; Sørlie et al., 2003; Parker et al., 2009). Other successful signatures have been developed with a different scope, for instance, to predict metastases development from primary tumor expression profiles (van't Veer et al., 2002; Wang et al., 2005).

These signatures suggest that there are different underlying molecular mechanisms in distinct breast cancer subgroups. However, when comparing the signatures established for the same prognostic purpose, they exhibit very little

overlap. This lack of agreement raises the question of whether the signature genes represent primary drivers of the disease or rather secondary downstream factors (Ein-Dor et al., 2005, 2006).

1.2.1 Wnt signaling in breast cancer

Aberrant regulation of Wnt signaling in general is critical for breast cancer initiation as well as progression (Howe and Brown, 2004). Smid et al. (2008) further demonstrated that molecular subtypes of breast cancer exhibit different levels of Wnt pathway activity. A number of important Wnt pathway players showed enrichment in the basal subtype and in all breast cancers, which later developed metastases.

The particular roles of distinct Wnt pathways in breast cancer context are still poorly understood. Altered expression of several non-canonical Wnt pathway members was associated with aggressive breast cancer subtypes. Higher expression of Ror2 and Ror1 receptors has been linked to breast cancer metastases (Klemm et al., 2011) as well as to shorter overall survival of breast cancer patients (Zhang et al., 2012; Henry et al., 2014). Furthermore, Wnt5a and Wnt5b were found to be upregulated in the basal-like MDA-MB-231 cells (Klemm et al., 2011). However, other studies reported contradicting evidence, with Wnt5a either enhancing or suppressing invasiveness of different breast cancer cell types (McDonald and Silver, 2009). Moreover, also canonical Wnt signaling via β-catenin has been implicated in the basal breast cancer subtype (Khramtsov et al., 2010; Yang et al., 2011).

Therefore, the currently available data offer rather ambiguous evidence on the activation of Wnt cascades and their key players in breast cancer. However, it is of high clinical interest to determine which particular Wnt pathways, including their domains and targets, are deregulated in poor prognosis breast cancer.

1.3 Gene expression analysis

The abundance of mRNA in a cell can now be quantified easily and reproducibly by gene expression profiling technologies such as microarray and the more recent

alternative: deep sequencing of RNA (RNA-Seq). These high-throughput measurements of RNA transcripts allow to determine the functions of the genes or their association to particular phenotype.

The most popular microarray chips such as Affymetrix or Agilent measure a single sample on one slide. These chips consist of thousands of short oligonucleotide probes spotted on solid substrate. The probes represent genomic DNA complementary to the specific transcript whose presence is to be investigated. First, RNA is extracted from the biological sample and labeled with biotin or a fluorescent tag. The labeled RNA is then hybridized to a microarray and washed off. Subsequently, the slide is scanned with laser light to quantify the fluorescent intensities. In contrast to microarrays, whose measurement range is limited to the spotted probes, RNA-seq allows for the complete annotation and quantification of all genes as well as their isoforms. With this approach the transcripts in the sample are directly sequenced and sequence reads are then aligned to a reference genome or transcriptome.To assess the gene expression levels the mapped reads are counted (Wang et al., 2009).

Within the framework of gene expression profiles, two main goals of analysis can be defined broadly: The first is to relate expression patterns with a certain phenotype (e.g. survival or disease relapse-free survival annotation) and the second is to identify differential genes between two conditions (Tarca et al., 2006). Depending on the desired goal, the experimental design and the choice of algorithms need to be adjusted.

The first category typically requires a bigger cohort of patients for which survival or other annotation data was collected along with expression profiles. The goal of this analysis is to discover new patient subgroups based on potential prognostic features. A popular application within this category is the use of hierarchical cluster analysis, which aims to discover patient subsets that share common expression patterns (Butte, 2002). Agglomerative clustering organizes expression profiles of the patients in an iterative manner, merging the two closest patients into a new composite object until all patients have been merged. As a measure of similarity correlation distance or euclidean distance can be utilized (Gibbons and Roth, 2002). The resulting hierarchical cluster tree is called a *dendrogram* and its branches correspond to individual patients. A next common step is to cut the dendrogram into clusters representing distinct

groups of patients, which can be further compared for differences in survival or disease progression.

In the second category two experimental groups are compared, such as control versus perturbation condition, or healthy versus disease state. The comparisons result in long lists of differentially expressed genes (DEGs) which are often challenging to interpret. Further analyses are usually performed to extract comprehensive information about the system under study. Hence, the results of differential analysis are placed into the context of existing literature sources. This can be achieved in an automatized fashion by integration with prior pathway and/or network knowledge.

1.4 Representation of pathway knowledge

Due to the complexity of transcriptomic data the results of a differential analysis are challenging to interpret. Prior biological knowledge can be used to link these results back to known molecular processes like signaling pathways. To this end, the pathway data needs to be represented in a computable format, such as a gene set or a graph model (Kitano et al., 2005).

The simplest graph representation, which captures pathway structure, depicts molecules as nodes and interactions as unweighted undirected edges. However, based on the character of signaling pathways a more natural representation is a directed graph (Schaefer, 2004). Within directed graphs the edges can be equally weighted ($w = \{1\}$) or the edge weights can distinguish inhibition and activation processes ($w = \{-1, 1\}$) between two nodes. Further edge annotation can include additional molecular events such as modifications, phosphorylation, translocation or transcription/expression. Moreover, in the concept of hypergraphs the edges are able to connect more then two nodes, capturing thus multiple functional and causal dependencies (Ritz et al., 2014). On the node level straightforward annotation depicts one node as one molecule, for instance a gene, a protein or a small molecule. More advanced models do not restrict nodes to one-to-one relationships but can include compound nodes with nested structure (Fukuda and Takagi, 2001) or metanodes to represent, for instance, protein complexes (Hu et al., 2007).

On the one hand, with the simpler representation some biological information is lost. But on the other hand, with more sophisticated models their inherent complexity can be problematic in terms of applicability and interpretability. Nevertheless, various models of a pathway can be used for integration with experimental data, provided they are in a computer-readable form.

1.5 Enrichment analysis

To annotate lists of DEGs enrichment analysis is a frequently used bioinformatic approach. There are many tests, which aim to detect pathways significantly altered between of two experimental conditions based on expression profiles (Khatri et al., 2012; Maciejewski, 2013). These tests have been implemented into a plethora of tools. The enrichment tools differ in many aspects ranging from the null hypothesis that is tested, through the statistical formulation, pathway data encoding and database support, up to the software implementation and utility.

Within the scope of the thesis two categories of enrichment methods are of interest, defined by the way in which pathway information is incorporated into the analysis. In the traditional enrichment approach a pathway is considered as a simple gene list omitting any knowledge of the gene and protein relations. Methods belonging to this category are referred to as gene-set (GS) analysis methods. Another way of integrating a pathway into enrichment analysis takes into account its topological structure. These methods are referred to as pathway topology-based (PT-based).

The particular methods of focus in this thesis are described in detail in the 2.2.1 *Gene-set enrichment methods* and 2.2.2 *Pathway topology-based methods* sections. Here I provide the reader with a general overview of the methodological principles within these two enrichment analysis approaches.

1.5.1 Gene-set enrichment approach

Initial tools for enrichment analysis fall into category of GS methods and most of the classification strategies and criteria are defined within this approach. Notably, the earlier GS methods, such as Onto-Express, GoMiner or

GoStat (Khatri et al., 2002; Zeeberg et al., 2003; Beißbarth and Speed, 2004), did not consider a gene set as a curated pathway gene list but as a Gene Ontology term (Ashburner et al., 2000).

In regard to the null hypothesis of the statistical tests they employ, the enrichment methods can be categorized into two groups: competitive and self-contained (Goeman and Bühlmann, 2007). Competitive methods compare genes in a pathway to its complement usually represented by the rest of the genes measured in the experiment. This approach is naturally linked with gene sampling for p-value calculation. Self-contained methods consider only genes within a pathway and test their association with the phenotype by subject sampling for significance assessment. Therefore, the competitive methods indicate whether there is a difference between the gene set and random gene sets of the same size in terms of association with phenotype, whereas the self-contained methods state how strong the association is, while not considering other gene sets at all. Both approaches have their limitations. On the one hand, competitive methods coupled with a gene sampling assume independence of genes, which is simply not true in the most cases. On the other hand, self-contained methods have been criticized for being too powerful and yielding too many significant gene sets. Furthermore, the number of experimental replicates is often too low for the purpose of subject sampling (Goeman and Bühlmann, 2007).

Another classification of enrichment methods separates them into over-representation analysis (ORA) and functional class scoring (FCS) groups (Khatri et al., 2012). ORA is the earliest strategy for enrichment analysis and is referring to 2×2 table methods such as Fisher's exact test, hypergeometric test and chi-squared test. It represents exclusively the competitive approach. The main drawback of ORA is that it requires a strict cut-off in the list of DEGs and that the enrichment results are strongly dependent on this chosen threshold. Therefore the FCS approach was suggested to overcome this difficulty. The FCS methods usually work in three steps: First genes are scored, then gene level scores are transformed into a pathway level score and finally the significance of the observed pathway level score is assessed. The FCS group includes both competitive and self-contained methods, depending on the pathway-level transformation and significance assessment of the pathway level score.

1.5.2 Pathway topology-based approach

One of the first PT-based methods was impact analysis introduced by Draghici et al. (2007). Since then this approach has become very popular, resulting in a number of various PT-based algorithms that have been published (Mitrea et al., 2013).

In contrast to traditional GS methods, the methodological concepts described in previous section have not been defined for the new PT-based group in such explicit terms. In many cases the concepts can be easily extended. For instance, if a PT-based method applies a strict threshold in the gene list it falls into ORA category. If in addition to topological information only expression data of the pathway genes are used to infer pathway significance, thereby omitting expression information of the genes outside the pathway, the approach reflects the self-contained concept. However, due to the inherent complexity of PT-based algorithms, in some cases it might be difficult to draw a strict line.

Considering the PT-based methods, an additional important categorization is based on how the topological information is exploited and incorporated into calculations. In regard to the extent of topological information, some methods consider the position of a gene in the entire pathway structure, e.g. by impact factor or betweenness measure, but some account only for close interaction partners termed neighbors (see 1.6.2 *Network analysis* and 2.2.2 *Pathway topology-based methods* sections for an explanation of the topological measures). Next, the topological information itself can be incorporated into an algorithm in various ways. The most straightforward approach is weighted GS analysis, where weights are assigned based on the topological measures (Gu et al., 2012). Further methods combine information from the standard ORA/FCS with a specific topology-based scoring system (Tarca et al., 2009; Dutta et al., 2012) or estimate pathway significance by using multivariate scoring models (Massa et al., 2010). For a detailed survey of PT-based methods with different pathway scoring systems the review of Mitrea et al. (2013) is recommended.

Within this thesis the term *enrichment analysis* is used as the most general label comprising methods of all categories. Usage of the term *pathway analysis* can be ambiguous: Sometimes it is used in the most general sense (Khatri et al., 2012), but sometimes it implies that the method accounts for the pathway graph

structure. Therefore, the latter is here referred to as *PT-based enrichment analysis* to avoid confusion.

1.6 Molecular networks

Molecular processes are typically regulated by coordinated effects of multiple interacting molecules. For instance, a defect in a particular gene does not affect only itself but also the activity of other genes and their products. Therefore, to determine underlying molecular processes of phenotypic or gene expression changes, it is helpful to study them in the view of molecular networks (Barabási et al., 2011). From the bioinformatic point of view, molecular networks are a representation of how genes and proteins cooperate in a given biological system. There are multiple types of molecular networks originating from different data types and focusing on different processes (see 1.6.1 *Types of networks*). On the one hand, it is possible to study a network on its own and the network architecture itself can reveal important functional principles and topological properties (see 1.6.2 *Network analysis*). On the other hand, a network can be integrated with experimental data in order to identify the part of the network that is affected in the experiment (see 1.6.3 *Network integration*).

1.6.1 Types of networks

Based on the type of interaction, several network types can be distinguished, which fall into two major categories – physical and functional interactions (Mitra et al., 2013).

The physical interaction group comprises protein-protein interaction (PPI), metabolic, regulatory, and signaling networks. A PPI network generally represents undirected interactions between pairs of binding proteins, usually detected by high-throughput yeast two-hybrid screens. Metabolic networks have varying representations: Nodes are associated with reactants and edges with reactions; however, the latter can also represent enzymes catalyzing these reactions. A regulatory network describes protein-DNA binding, which represents translational regulation. In these, two types of nodes are typically connected with directed edges – reflecting transcription factors that regulate target gene ex-

pression. Signaling or signal-transduction networks are usually defined less explicitly. They can be described as interconnected chains of post-translational modifications and other biochemical reactions, PPIs and/or changes in gene expression (Albert, 2005; Choudhary and Mann, 2010). Therefore, a signaling network can be seen as an hybrid involving several physical interaction types. For a comprehensive overview of these and other physical interaction network types which are not discussed in this thesis (e.g.: RNA-RNA, drug-target interactions) see reviews of Albert (2005), Barabási et al. (2011) and Vidal et al. (2011).

The second category of networks involves functional interactions, such as gene-gene or gene-drug interactions. Genetic or so-called epistatic interaction reports on interaction between two mutations when the combination of mutations results in a different phenotype than expected from the phenotypes of each mutation individually. Moreover, also co-expression networks depicting correlated expression between genes fall into the group of functional gene-gene interactions. The second functional interaction type, the gene-drug interaction, can be seen as an equivalent of genetic interaction in a sense that a gene perturbed in the presence of a drug results in a combined effect more or less severe than expected (Ryan et al., 2013).

Within this thesis the major focus is on signaling networks. However, all these networks represent a complementary, although, rather simplified view on the complex cellular system. Nevertheless, this simplification enables us to investigate inherent properties of the networks and integrate the networks with different molecular profiles.

1.6.2 Network analysis

Within the network analysis graph theory concepts are utilized to describe the structure of a given network. To elucidate the topological structure, properties such as size of the network, node features and network communities are characterized.

One of the basic characteristics of a node in a network is its degree, k, representing the number of interaction partners. In a directed network in- and out-degree can be distinguished by summing up the numbers of incoming

and outgoing edges, respectively. Furthermore, the degree distribution for the entire network can be defined as the probability that a given node has exactly k edges. Many molecular networks are considered to be scale-free meaning their degree distribution follows a power law (Barabási and Albert, 1999). It implies that there is only a relatively small number of nodes that are highly connected, whereas the most of the nodes have a low degree. The nodes with many interaction partners are often called hubs and their roles have been intensively studied in model organisms as well as humans. Hypotheses have been formulated that the hubs are encoded by essential genes that hold the network together, whereas nonessential, disease-related genes are typically not represented by hubs (Goh et al., 2007).

Another node characteristic is termed betweenness centrality. It is defined as the fraction of the shortest paths in a directed graph passing through a given node out of the shortest paths between all node pairs. High betweenness nodes are also called *bottlenecks* and have the tendency to correlate with essentiality (Yu et al., 2007). A corresponding measure can also be defined on edges as the fraction of shortest paths between all pairs of nodes that pass through the given edges out of all shortest paths.

Further network concept defines the nodes closely surrounding a certain node as its neighborhood. The first order neighborhood around a given node includes its directly connected interaction partners – neighbors. Second order neighborhood comprises neighbors that are not farther away from the node then two steps, and so forth.

Another network analysis approach aims to find modules that represent highly interconnected local regions in network topology. Such clusters or so-called communities have dense connections between the cluster nodes and spare connections between nodes of different clusters. Several algorithms utilizing different quantifying measures were proposed. In their seminal work Girvan and Newman (2002) proposed a divisive algorithm based on progressive edge removal. The edge to be removed is chosen based on the highest edge betweenness score, which is recalculated after each removal. The idea is that by removing "between-communities" edges, the network splits into its natural communities. Later on Newman and Girvan (2004) suggested a modularity as property of network which can be used as cluster criterion for the network

division into communities. The modularity is defined as difference of the number of within-community edges and the expected number in an equivalent network with edges distributed at random. In order to find optimal network modularity an optimization algorithm is employed (Clauset et al., 2004). The algorithm starts with each node as its own community and repeatedly merges pairs of communities whose merge results in the greatest modularity increase, until the point when further merging only reduces modularity. By maximizing modularity the best division of the network is found.

Community as well as *neighborhood* are topological modules, which do not account for any function of the gene nodes. Nevertheless, studies have suggested that components within a topological module can have similar or related functions (Barabási et al., 2011). Further types of modules which go beyond the topology information can be identified by network integration with different data types.

1.6.3 Network integration

Generally, network integration comprises two distinct approaches: The integration of two or more types of networks and the integration of a network with experimental data. Here, the focus is on network integration with experimental data, specifically with transcriptomic profiles. On the one hand, the integration facilitates the analysis and interpretation of comprehensive gene expression profiles. On the other hand, it aims towards extraction of differentially expressed subnetworks (or so-called active or responsive functional modules) from networks. These context dependent subnetworks then help to elucidate underlying molecular changes in a biological system responding to a perturbation or a stimulus (Wu et al., 2009; Mitra et al., 2013).

To extract such a subnetwork from a large scale network usually three steps are required: The first step is scoring the network nodes based on some measure representing differential expression. The second is aggregating the scores over all nodes in each subnetwork, and the third is finding the "best" score subnetwork. One of the first methods within this field was work of Ideker et al. (2002) applying a simulated annealing algorithm to search a high score subnetwork. Later on also exact solutions were suggested, for instance by Dittrich et al. (2008). They transformed the problem of finding

the maximum score subnetwork into the prize-collecting Steiner tree problem and used integer-linear programming to find the solution. In more simple scenarios with unweighted network nodes the classical Steiner tree problem can be employed to identify a minimal size subnetwork containing all nodes of interest (Sadeghi and Fröhlich, 2013).

An add-on to these integration approaches goes one step further and uses identified subnetworks as discriminative or prognostic markers to predict patient outcome. For instance, Chuang et al. (2007) showed that subnetwork markers achieved higher accuracy in the classification of breast tumors than individual genes without network information. However, other studies reported that equal or even better classification can be achieved using randomized networks (Lavi et al., 2012; Staiger et al., 2012).

Nevertheless, the integration of expression profiles with networks provides great potential for identifying phenotype associated genes as well as markers for disease prognosis.

1.7 Aims and organization of the thesis

Two principal questions define the focus of this thesis and the particular aims are derived from these questions:

- What are benefits and costs of integrating pathway-topology information into enrichment analysis?
- Which module of the Wnt signaling network is active in aggressive breast cancer?

These questions share ideas and methodological overlap in terms of processing signaling pathway knowledge, applying graph theory concepts, and integrating topology information with experimental data.

Within the enrichment analysis framework, several studies have already compared various GS analysis methods (Abatangelo et al., 2009; Evangelou et al., 2012; Tarca et al., 2013). However, only little effort has been dedicated to the comparison of PT-based methods and evaluation was limited to a small number of real datasets (Jaakkola and Elo, 2015). Therefore, the first aim related to the first question is to:

1. Comparatively evaluate different enrichment analysis approaches: traditional gene-set methods versus pathway topology-based methods.

I aim to compare three GS and four PT-based methods using simulated gene expression data as well as a collection of 24 benchmark datasets. Comparison of these approaches involves two major challenges. First, simulation of suitable expression data for the PT-based method evaluation is a rather comprehensive problem, as the structure of deregulated pathways has to be reflected in the synthetic data. Therefore, a complex simulation scheme is described in the separate 2.3 *Simulations* section within the 2 *Materials and Methods* chapter. Second, the pathway models differ between the methods – not only the obvious gene sets versus pathway graphs – but particularly different graph representations used within distinct PT-based methods. In order to perform a fair comparison unified pathway input is needed. Thus, I aim to integrate the same pathway data customized for each evaluated enrichment method. The processing of pathway information is presented in the 2.2.3 *Parsing pathway knowledge from databases* section. The resulted parsed pathways are described in the 3.1 *Parsed pathways* section, whereas the results of the comparative evaluation of the methods are presented in the 3.2 *Comparison of enrichment methods* section.

In regard to the second question, Figure 1.1 delineates a simplified conceptual workflow of multiple steps leading to the answer.

Within the Wnt signaling field the separation of canonical and non-canonical Wnt pathways proved useful for a better understanding of these processes. However, this separation is often not well distinguished in pathway databases. Reliable models of the Wnt pathways, which are suitable for further bioinformatic analysis, are needed. Hence, the next aim can be formulated as follows:

2. To assemble public Wnt pathway knowledge into signaling network models representing distinct Wnt pathways.

This is a four-fold task as the pathway information has to be collected, parsed, curated, and finally merged to create network models (Figure 1.1A-B). Notably, the data parsing step largely overlaps with the generation of

pathway input for the enrichment methods, therefore it is presented together in the 2.2.3 *Parsing pathway knowledge from databases* section. The final models represented as signaling networks and their graph properties are described in the 3.3 *Wnt networks* section.

Such Wnt models allow further investigation of the underlying biological processes when integrated with experimental data. In particular, I am interested in elucidating the role of canonical and non-canonical Wnt signaling in breast cancer. To study pathway activation in this context, the estrogen receptor positive MCF-7 breast cancer cell line is utilized as a model system. The phenotype of the MCF-7 cell line without any intervention experiment is considered to correspond to lumA breast cancer subtype with favorable clinical prognosis. The literature is rather ambiguous on the role of Wnt signaling in breast cancer. Based on the results of my collaborators (Klemm et al., 2011) the working hypothesis within this thesis is that activation of the non-canonical Wnt pathway stimulates cell proliferation and migration. Following this hypothesis, cell invasiveness of MCF-7 can be enhanced by perturbations of the non-canonical Wnt pathway members the (Figure 1.1C) and subsequently the targets of the perturbations can be identified by gene expression profiling (Figure 1.1D). I aim to integrate such expression data with the newly constructed Wnt networks in order to identify specific Wnt activation modules (Figure 1.1E). Therefore, the next major objective of this thesis can be summarized as:

3. To identify an expression-responsive module within the Wnt networks relevant for invasiveness of breast cancer cells.

Since generation and expression profiling of the cell lines are carried out by collaborators[(1)], I only briefly summarize these experiments in the *Materials* section 2.1.2 *Newly generated RNA-Seq data*. The results of differential and enrichment analyses as well as signaling network integration are described in detail in the *Results* sections 3.4 *Sequenced cell line and perturbation targets* and 3.5.1 *Non-canonical Wnt module*. Furthermore, perturbations of the cells are expected to have impact also on the genes outside the Wnt signaling. Thus, I aim to further explore the expression data by utilizing a PPI network. The corresponding results are described in the 3.5.2 *PPI network and hubs* section.

[(1)]Specific contributions of collaborators are stated at appropriate places in the text.

Finally, I seek to validate the results of the network integration in the clinical context of breast cancer patients. I aim to evaluate the relevance of gene signatures originating from the results of signaling and PPI networks integration in respect to metastasis-free survival (MFS) prognosis (Figure 1.1F-G). Moreover, I am interested in detecting individual genes with prognostic potential in this context. Hence, the last aim is defined as:

4. To evaluate the prognostic power of the gene signatures in terms of MFS and to identify potential prognostic markers.

The results of this analyses are reported in the 3.6 *Breast cancer metastasis-free survival study* section.

In the 4 *Discussion* chapter I first discuss challenges which arose within the parsing of signaling pathway data and their representation for further analyses (section 4.1 *Parsing and representing pathway knowledge*). Then I critically evaluate enrichment methods and further elaborate on issues dealing with general concepts within the enrichment analysis framework (section 4.2 *Enrichment methods*). In the 4.3 *Wnt networks* section I discuss reliability as well as limitations of the newly constructed Wnt models. The differential targets of experimental perturbations and their integration with networks are discussed in the 4.4 *Targets and network integration* section. The final *Discussion* section 4.5 *Breast cancer metastasis and prognostic genes* discusses the results of the patient cohort study and potential prognostic impact of the new gene signatures.

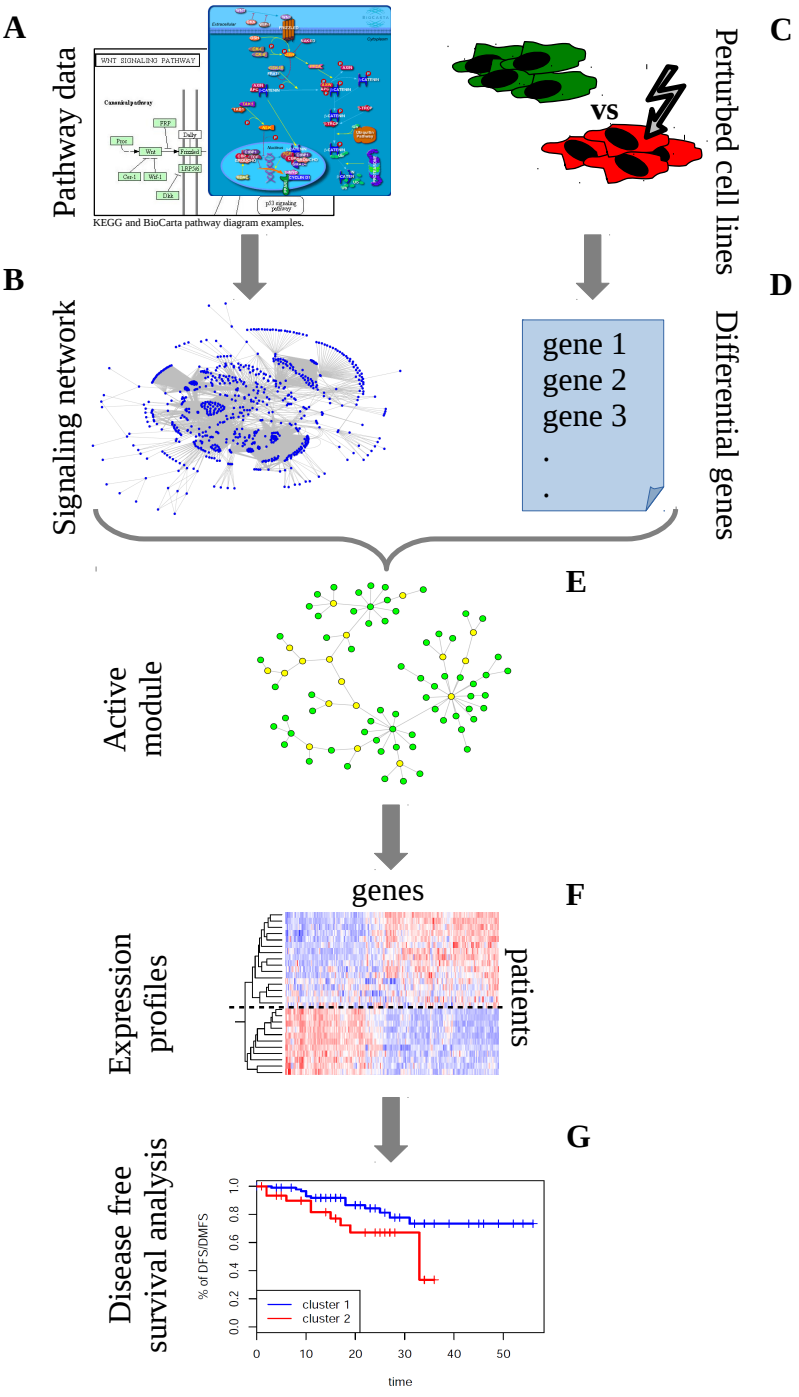

Figure 1.1. *Conceptual workflow: Signaling network (A-B) and gene expression data (C-D) integration results in an active module identification (E). The results are further used towards translation into metastasis-free survival study of breast cancer patient cohort (F-G).*

Materials and Methods

This chapter is structured into three sections. 2.1 *Materials* section summarizes gene expression, pathway and network data utilized and generated within the scope of this thesis. 2.2 *Methods* provides the reader with a comprehensive insight into employed approaches and algorithms. To evaluate different enrichment methods, also synthetic expression data were generated using multiple simulation scenarios with a number of parameters being explored. Therefore, the complex simulation settings are described in an individual section – 2.3 *Simulations*.

2.1 Materials

The section provides brief overview of the public experimental datasets (2.1.1) which were used for the analyses. In case of newly sequenced expression data (2.1.2), I summarize how the cell lines were treated and the libraries prepared. Finally, last subsection (2.1.3) describes databases and repositories from which prior information on pathways and networks was retrieved.

2.1.1 Public microarray data

Two large compendium datasets of expression profiles were analyzed within this work: breast cancer patient data and benchmark data comprising multiple diseases.

2.1.1.1 Patient datasets

The breast cancer patient data is a collection of ten microarray datasets hybridized on Affymetrix Human Genome HG-U133 Plus 2.0 and HG-U133A

arrays. The datasets were retrieved from the Gene Expression Omnibus (GEO) (Barrett and Edgar, 2006) data repository under the accession numbers GSE25066, GSE20685, GSE19615, GSE17907, GSE16446, GSE17705, GSE2603, GSE11121, GSE7390, and GSE6532. Each dataset was processed using RMA algorithm (Irizarry et al., 2003) and only samples with full metastasis free survival annotation (meaning annotated with both the metastasis/distant-relapse event and the follow-up time information) were selected. The datasets were combined on the bases of HG-U133A array probe IDs into a single expression matrix and quantile normalized. Metastasis free survival annotation was compiled into the same time unit (years) for all samples. Within further processing steps of the data, the breast cancer molecular subtypes were predicted. A single sample predictor was fitted for each patient using *pam50* intrinsic genes list (Parker et al., 2009) as implemented in *genefu* R-package (Haibe-Kains et al., 2012, 2011). Particular molecular subtype was assigned to a patient when prediction strength > 0.5.

2.1.1.2 Benchmark datasets

For the purpose of enrichment methods comparison, 24 datasets from the *KEGGdzPathwaysGEO* R-package (Tarca et al., 2012) were used as benchmark data(Table 2.1). Disease datasets comprise 880 samples representing 12 distinct diseases and corresponding controls. Each of the 24 datasets was matched with the corresponding KEGG pathway according to its name, e.g. a dataset of colon cancer patients was associated with the colorectal cancer pathway. Such a pathway was then called a target pathway and its p-value and rank in the database were further evaluated (Tarca et al., 2012; Evangelou et al., 2012).

2.1.2 Newly generated RNA-Seq data

The human breast cancer cell line MCF-7 was obtained from the American Type Culture Collection (ATCC, Rockville, USA) and was cultured in RPMI-1640 media (PAA, Cölbe, Germany) supplemented with 10% fetal bovine serum (FCS; Sigma, Munich, Germany). For Ror2 over-expression, the plasmids pcDNA 3.1/Zeo(+) (Invitrogen, Paisley, UK) and pcDNAhsROR2 were introduced into MCF-7 cells using the Nanofectin transfection reagent (PAA, Cölbe,

Germany). Stable expression was achieved by selecting for zeomycin (100 µg/ml) resistance. For stimulation experiments the cells were treated for 24h with Wnt5a (100 ng/ml, R&D systems) prior to cell lysis. RNA was isolated using Trizol reagent, including a DNase I (Roche, Mannheim, Germany) digestion step. All cell line cultures and intervention experiments were carried out in the Binder/Pukrop lab[(1)] by Dr. med. Florian Klemm, Dr. med. Annalen Bleckmann and Dr. Kerstin Menck.

The cell lines were further sequenced at TAL (*Transkriptomanalyselabor*) by Dr. Gabriela Salinas-Riester[(2)]. Library preparation for RNA-Seq was performed using the TruSeq Stranded Total RNA Sample Preparation Kit (Illumina, RS-122–2201) starting from 1000 ng of total RNA. Accurate quantitation of cDNA libraries was performed using the QuantiFluor TM dsDNA System (Promega). The size range of final cDNA libraries was determined applying the SS-NGS-Fragment 1-6000 bp Kit on the Fragment Analyzer from Advanced Analytical (320 bp). cDNA libraries were amplified and sequenced by using the cBot and the HiSeq2000 from Illumina. Sequence images were transformed with Illumina software BaseCaller to bcl files, which were demultiplexed to FastQC files with CASAVA v1.8.2. RNA-Seq data were uploaded to the GEO repository under the accession number GSE74383.

The invasive capacity of the MCF-7 cells was measured in a modified Boyden chamber as previously published by Hagemann et al. (2004). Cells were seeded in triplicates onto an ECM-coated (R&D systems) polycarbonate membrane (pore diameter: 10 µm, Nucleopore), optionally stimulated with Wnt5a (400 ng/ml, R&D systems) and incubated for 96 h at 37°C. The number of invasive cells in the lower wells was counted and related to the unstimulated control. All invasion assays were carried out in three biologically independent experiments by Dr. Kerstin Menck and Dr. Matthias Schulz[(3)].

[(1)] Department of Hematology and Medical Oncology, University Medical Center Göttingen.
[(2)] DNA Microarray and Deep-Sequencing Facility Göttingen, Department of Developmental Biochemistry, University of Göttingen
[(3)] Department of Hematology and Medical Oncology, University Medical Center Göttingen.

2.1.3 Pathway and network databases

A number of public databases systematically collect and curate signaling pathway information. My main focus was on the pathway databases that store their data in the Biological Pathways Exchange (BioPAX) format. BioPAX is a standard Web Ontology Language (OWL)-based model encoding the pathway knowledge at the molecular level. Pathway databases with BioPAX export utilized in this work include BioCarta (Nishimura, 2001), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2004), Reactome (Croft et al., 2011), Pathway Interaction Database (PID) (Schaefer et al., 2009) and a meta-database Pathway Commons (Cerami et al., 2011).

For enrichment methods evaluation the pathways from KEGG were downloaded in BioPAX level 3 export on March 2013 and only non-metabolic pathways were selected. Five pathway databases (Table 2.2) were scanned in order to collect publicly available Wnt signaling data. The database exports of BioPAX level 3 files were downloaded in March 2014. Pathways of interest were selected according to the presence of important Wnt signaling components. Further pathways processing and stratification are described in the 2.2.3 *Parsing pathway knowledge from databases* section.

For protein-protein interaction (PPI) network integration the BioGRID (Stark et al., 2006) database was utilized. The BioGRID interactome was downloaded at September 2014 (3.2.116) as Tab 2.0 Delimited Text file for *Homo sapiens*. Only protein-protein interactions were selected, omitting genetic interactions.

GEO accession	Disease/Target pathway	Samples
GSE781	Renal cell carcinoma	17
GSE1297	Alzheimer's disease	16
GSE3467	Thyroid cancer	18
GSE3585	Dilated cardiomyopathy	12
GSE3678	Thyroid cancer	14
GSE4107	Colorectal cancer	22
GSE5281_EC	Alzheimer's disease	21
GSE5281_HIP	Alzheimer's disease	23
GSE5281_VCX	Alzheimer's disease	31
GSE6956AA	Prostate cancer	10
GSE6956C	Prostate cancer	16
GSE8671	Colorectal cancer	64
GSE8762	Huntington's disease	22
GSE9348	Colorectal cancer	82
GSE9476	Acute myeloid leukemia	63
GSE14762	Renal cell carcinoma	21
GSE15471	Pancreatic cancer	70
GSE16515	Pancreatic cancer	30
GSE18842	Non-small cell lung cancer	88
GSE19188	Non-small cell lung cancer	153
GSE19728	Glioma	21
GSE20153	Parkinson's disease	16
GSE20291	Parkinson's disease	33
GSE21354	Glioma	17
Total:		**880**

Table 2.1. *Summary of 24 benchmark datasets from the KEGGdzPathwaysGEO R-package. The columns represent the accession number from GEO database, the name of target pathway and the number of samples for each dataset.*

Database	N
BioCarta	4
KEGG	3
Pathway Commons	2
PID	7
Reactome	9

Table 2.2. *Databases from which Wnt signaling data was retrieved with the number of pathways (N) used for network construction.*

2.2 Methods

Three gene-set (GS) and four pathway topology-based (PT-based) enrichment methods, which were comparatively evaluated in this study, are described in detail in 2.2.1 and 2.2.2 sections and summarized in the Table 2.3. Further, in 2.2.3 section the workflow of parsing and editing pathway data from public databases is defined. Finally, within the last three parts I briefly summarize differential analysis of expression data (see 2.2.4), network analysis and network integration approaches (see 2.2.5), and survival analysis (see 2.2.6).

Within this thesis the majority of bioinformatic as well as statistical analyses were performed in the environment for statistical computing R (Team, 2012).

Method name	Approach	ORA/FCS	R-function/package	Null hypothesis
Wilcoxon rank sum	GS	FCS	wilcox.test	Competitive
Kolmogorov-Smirnov	GS	FCS	ks.test	Competitive
Fisher's exact	GS	ORA	fisher.test	Competitive
SPIA	PT-based	ORA-like	SPIA 2.12.0.	Competitive
CePa ORA	PT-based	ORA-like	CePa 0.5.	Competitive
CePa GSA	PT-based	FCS-like	CePa 0.5.	Self-contained
PathNet	PT-based	unclass.	PathNet 1.3.0.	Competitive

Table 2.3. *Summary of the gene set (GS) enrichment and the pathway topology-based (PT-based) methods evaluated in this study. The seven methods were stratified into over-representation analysis (ORA) and functional class scoring (FCS). However, for PT-based methods this classification is not always explicit, therefore '-like' suffix is used and 'unclass.' for an unclassified method. Next, the utilized R-functions for GS methods and R-packages of PT-based methods are stated. According to the null hypothesis most of the methods are competitive, only the GSA variant of the CePa method is self-contained.*

2.2.1 Gene-set enrichment methods

Basic statistical tests were chosen to represent the GS analysis approach: Wilcoxon rank sum (WRS), Kolmogorov-Smirnov (KS) and Fisher's exact (FE) tests. These tests were implemented in various flavors and extensions in multiple tools and software packages (Khatri et al., 2002; Mootha et al., 2003; Beißbarth and Speed, 2004; Barry et al., 2008). The R-functions *wilcox.test*, *ks.test* and *fisher.test* were utilized to perform WRS, KS and FE analysis, respectively. All three tests are competitive gene sets approaches, which require a list of p-values for differential expression (Beissbarth, 2006). WRS and KS are functional class

scoring (FCS) methods transforming the list of p-values into ranks. WRS tests whether the distribution of ranks of the genes in a set is shifted to the left from a distribution of ranks of the genes in corresponding complement to the gene set. KS test compares the ranks of genes in a set to the uniform distribution. The FE test is based on over-representation analysis (ORA). Therefore, a cut-off in the list of differentially expressed genes (DEGs) needs to be defined: Here, false discovery rate (FDR) below 0.05 ($FDR < 0.05$) was considered as the threshold. FE is testing independence of rows and columns in 2×2 contingency table (while the margins are fixed) and p-values are directly obtained using hypergeometric distribution.

2.2.2 Pathway topology-based methods

The evaluated PT-based algorithms come from three R-packages: *SPIA*, *CePa* and *Pathnet*. Two variants of the CePa method were implemented in the *CePa* R-package, the so-called CePa ORA and CePa GSA, which I further consider as two distinct methods.

SPIA (signaling pathway impact analysis) is an enrichment method which combines two types of evidence represented by two p-values (Tarca et al., 2009). The first p-value originates from a simple ORA, assuming that the number of DEGs in a given pathway follows hypergeometric distribution. The second, so-called perturbation p-value is computed in several steps and incorporates information on pathway topology. In order to obtain the perturbation p-value, first, for each gene g_i in a pathway a perturbation factor is computed:

$$PF(g_i) = \Delta E(g_i) + \sum_{j=1}^{n} \beta_{ij} \frac{PF(g_j)}{N_{ds}(g_j)}$$

where first term $\Delta E(g_i)$ captures the logarithm of the fold-change of a gene g_i and the second term describes the sum of perturbation factors of the direct upstream genes of a gene g_i normalized by the number of all downstream genes $N_{ds}(g_j)$. Each term of the sum is weighed by β_{ij} quantifying the type of interaction between the two genes: 1 and -1 for activation and inhibition, respectively. This results in an upstream gene influencing perturbation factors of many downstream genes. In the second step the perturbation accumulation

$Acc(g_i)$ at the level of each pathway gene is calculated. It is defined as the difference between the gene perturbation factor and its observed logFC:

$$Acc(g_i) = PF(g_i) - \Delta E(g_i)$$

Finally, the total pathway accumulated perturbation t_A is computed as a sum of the accumulated perturbations of pathway's genes:

$$t_A = \sum_i Acc(g_i)$$

Significance is assessed in a bootstrap procedure, resulting in a perturbation p-value. The two p-values are then combined into a global p-value for each pathway using Fisher's product test.

CePa (Gu et al., 2012) is a weighed gene set analysis approach in which weights are assessed by network centralities. The CePa ORA, first CePa method variant, weighs the nodes of DEGs according to one of five centrality measures and then sums them up to the pathway level score:

$$s = \sum_{i=1}^{n} w_i d_i$$

where d_i stands for binary variable identifying whether pathway node is differentially affected and w_i is the centrality of the node. To obtain a p-value the null distribution of the pathway score is generated by permuting the DEGs on a given pathway topology. There are five options to assess the node centrality: in- and out-degree, betweenness and in- and out-largest reach. Degree centrality measures the number of incoming or outgoing edges of the given node. Betweenness reflects the number of information streams passing through a given node. Largest reach quantifies how far a node can send or receive information. Along five centrality options CePa also calculates an equal weight model where all weights are set up to 1. Therefore, for each pathway six p-values are calculated and the authors of method recommend to try every centrality option in the search for significant pathways. Hence, for the purpose of method evaluation (see section 3.2) the smallest out of 6 p-values was selected to represent pathway significance. Furthermore, the CePa method proposed a node-based instead of a gene-based ID mapping approach. That means, if any member of a complex

or a group of genes residing in one node is differentially expressed then the node is considered as differentially expressed. Also nodes representing non-gene components of a pathway such as microRNA and small molecules are retained in the pathway topology. However, these last two features of the CePa algorithm were suppressed by using customized pathway input data (see section 2.2.3).

The second method variant, so-called CePa GSA performs self-contained univariate gene-set analysis. Firstly, the node level statistic vector $\mathbf{d}$ is weighted by centrality measures $\mathbf{w}$ and then they are transformed into pathway level statistics using transformation function f:

$$s = f(\mathbf{w} \cdot \mathbf{d})$$

resulting in pathway score s. CePa GSA implements several alternatives for both, the node level statistic and transformation function. The utilized default option for node level scores was the absolute value of t-statistic and for computing pathway level statistics the mean was used as transformation function. Then the significance of each pathway is assessed by permuting sample class labels.

PathNet (Dutta et al., 2012) is an enrichment method also combining two types of evidences, similarly to SPIA method. However, in contrast to SPIA, which combines two p-values on pathway-level, PathNet operates with gene-level p-values. The method considers so-called direct and indirect evidence. Direct evidence coming from expression data is represented by nominal p-values (p^D) of the DEGs. Indirect evidence of a gene is calculated from direct evidence p-values of all its neighbors in a pooled pathway. The pooled pathway is a big network created by merging all pathways presented in a given database. To calculate indirect evidence, first, indirect evidence score SI of a gene i is defined as

$$SI_i = \sum_{j \in G, i \neq j} A_{ij} \cdot (-log_{10}(p_j^D))$$

where G represents all genes in the pooled pathway and A denotes the adjacency matrix corresponding to the pooled pathway network. Secondly, the null distribution of the SI score is reconstructed by randomizing direct evidence p-values on the pooled pathway with a fixed topology and the corresponding indirect evidence p-value is estimated. Finally, a p-value for each gene is

obtained by aggregating direct and indirect evidence p-values using Fisher's method. The final pathway significance is assessed via a hypergeometric test.

2.2.3 Parsing pathway knowledge from databases

Signaling pathway data can be incorporated into bioinformatic analyses in a form of gene sets or signaling graphs, and the latter is certainly more challenging task. In order to integrate topological information as prior knowledge, the data have to be retrieved and handled in an appropriate way depending on the particular analysis approach. In this work I collected and processed pathway data for two main purposes: **(1)** To provide suitable pathway input for the enrichment methods and **(2)** to assemble multiple Wnt signaling pathways into Wnt networks. For both tasks multiple common steps were taken (Figure 2.1 A-B), but eventually also unique steps were performed for shaping input for enrichment analysis (Figure 2.1 C1) and to create Wnt network models (Figure 2.1 C2).

First, exports of BioPAX models were downloaded from the pathway databases (for the database list see section 2.1.3). The pathways in the BioPAX files were parsed into R using the *rBiopaxParser* R-package (Kramer et al., 2013) and represented as interaction graphs (Figure 2.1 A), in which directed edges denote activation or inhibition processes between the nodes. The pathway databases use different identifications to annotate pathway molecules. Different annotations were converted into HUGO Gene Nomenclature Committee (HGNC) gene symbols. After HGNC IDs on the graph nodes several editing steps had to be taken (Figure 2.1 B). First, in case that several nodes were annotated with the same gene symbol, these nodes were merged into a node, which shared all incoming and outgoing edges of the original nodes. Next, gene families or protein complexes often occupied a single node, which resulted in multiple symbol IDs embedded in the node. Such a node was split into multiple nodes and each one was assigned with a single gene symbol. Finally, non-gene pathway entities, such as small molecules, DNA, RNA, or nested pathways, were filtered out, while connectivity of the graphs was preserved by introducing new edges connecting the genes that were indirectly connected. Further processing steps differed between **(1)** creating input for enrichment methods and **(2)** assembling Wnt networks.

(1) Suitable pathway input for each enrichment method had to be provided implying graphs transformations (Figure 2.1 C1). For GS methods the pathway graphs were simply converted into lists of genes. However, PT-based methods required specific pathway topology inputs. To create SPIA pathway input, a graph of each pathway was transformed into a list of 2 adjacency matrices for activation and inhibition processes. Accordingly, a vector of weights β was set to $\beta = \{1, -1\}$ to reflect activation and inhibition, respectively. For CePa a pathway catalog was constructed comprising a list of pathways with the interaction IDs, and a table with the interaction IDs and corresponding input and output interaction components, and a mapping table. Input for PathNet consisted of an adjacency matrix of a pooled pathway, which was created by merging all pathways from KEGG database (see section 2.1.3), an interaction table with pathway IDs, and a mapping table.

All pathway data inputs were regenerated for the second simulation study (see Simulations section 2.3). In study 2 the pathway graph nodes were relabeled with new synthetic IDs to construct non-overlapping pathways with unique components, whereas the topology of the pathway remained intact.

(2) To create Wnt network models the parsed pathways were curated based on the literature and expert knowledge[4] into four groups representing: canonical Wnt signaling, non-canonical Wnt signaling, inhibition of canonical Wnt signaling, and regulation of Wnt signaling. Pathways which were considered to be too unspecific or general to be classified were discarded. In order to merge these pathways into signaling networks (Figure 2.1 C2), each pathway graph was transformed into the simple interaction format. Nodes without any interactions were excluded and the interaction tables were concatenated according to the assigned groups. Duplicated interactions were removed and the four interaction tables were finally transformed back into graph objects.

[4] Consulting with Dr. med. Florian Klemm and Dr. med. Annalen Bleckmann, Department of Hematology and Medical Oncology, University Medical Center Göttingen.

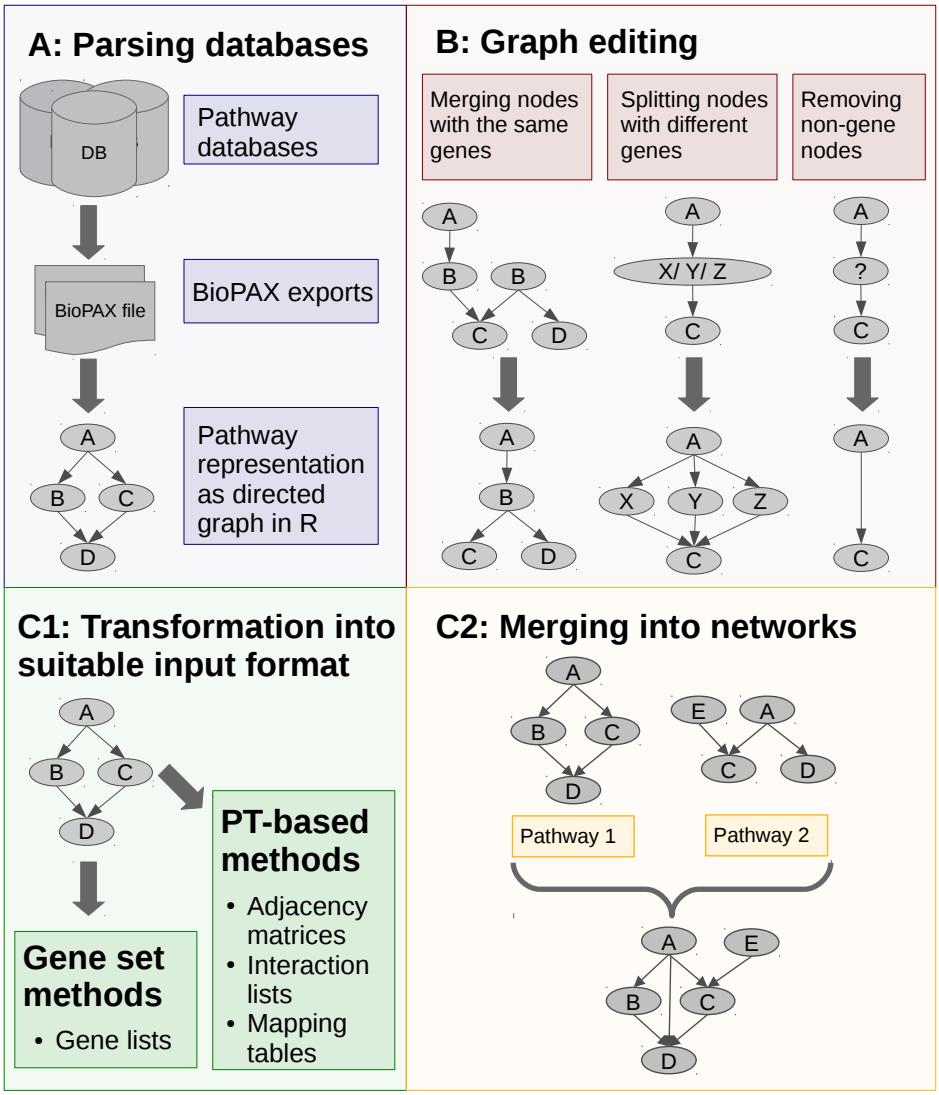

Figure 2.1. *(A) Database pathways in BioPAX format were parsed into R and represented as directed interaction graphs. (B) The graphs were edited in order to affiliate a single graph node with a single gene symbol ID. (C1) The adjusted graphs were either transformed into an appropriate input format for a particular enrichment method, (C2) or were merged into signaling networks to built Wnt models. When merging two pathways, the resulting graph contains all nodes and edges from both original graphs.*

2.2.4 Differential analysis

Differential expression analysis of microarray data was performed by fitting linear models using the empirical Bayes method as implemented in the *limma* R-package (Smyth, 2005). Within preprocessing steps prior the differential analysis, multiple probes per Entrez Gene ID were removed by retaining the

probe with the highest average expression. For the purpose of consequent enrichment analysis, the genes that could not be mapped onto any pathway were filtered out.

RNA-Seq data were 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). Further preprocessing steps were done using *edgeR* (Robinson et al., 2010) R package: Low expressed genes were filtered out by keeping the genes with at least one count-per-million read in at least three samples and read counts were normalized to the library size. Finally, differential genes between distinct conditions were identified by fitting negative binomial generalized linear models (Robinson and Smyth, 2008).

The p-values coming from the analyses of both microarray and RNA-seq platforms were adjusted for multiple testing using the method of Benjamini and Hochberg (1995) resulting in FDR and genes with $FDR < 0.05$ were considered significantly differentially expressed.

2.2.5 Network-based analyses

To explore network properties the central nodes were identified using two network centrality measures: Betweenness centrality implemented in the *betweenness* R-function and degree centrality implemented in the *degree* R-function from the *igraph* R-package (Csardi and Nepusz, 2006). The overlap nodes between the nodes scored best by both measures are referred to as key nodes in the signaling network (overlap of top 15) and as hub nodes in the PPI network (overlap of top 50).

To determine community structure in a graph, dense subgraphs were sought using the fast greedy modularity optimization algorithm implemented in the *fastgreedy.community* R-function (Clauset et al., 2004).

In the integration analysis of the network with expression data the differentially regulated targets were mapped onto the network nodes. Induced nodes were used as terminal nodes for the Steiner tree analysis based on shortest path approximation as implemented in the *SteinerNet* R-package (Sadeghi

and Fröhlich, 2013). A minimal spanning tree which contains all terminal nodes was searched in the undirected version of the network. In this analysis so-called Steiner nodes are introduced to ensure the connectivity of the tree. Resulting tree nodes were used to extract a differential subnetwork containing all original directed edges. For visualization purposes Fruchterman-Reingold layout (Fruchterman and Reingold, 1991) was utilized and the range for color coding of nodes was limited to ±2 fold-change.

2.2.6 Clustering and survival analysis

For the purpose of breast cancer patient data analysis the complete-linkage hierarchical clustering was performed based on Pearson correlation as the distance measure. When multiple probes corresponded to a single gene, a probe with highest average expression level for the gene was kept for the clustering analysis. To identify clusters representing different sample groups within the dendrogram the dynamic hybrid cut algorithm was utilized implemented in R-function *cutreeDynamic* from the *dynamicTreeCut* R-package (Langfelder et al., 2008). The algorithm detects the clusters in a bottom-up manner based on the dendrogram shape information and the correlation dissimilarity information among the patients. The minimum cluster size parameter was set as 12.5% of the patients when the whole dataset was clustered and 25% of the patients when only patients of a particular breast cancer subtype were clustered. Resulting clusters were subjected to a Kaplan-Meier (KM) analysis of metastasis-free survival (MFS) and KM curves were compared using a log-rank test implemented in *survival* R-package. When plotting the KM curves the first 15 years were visualized. The Cox proportional hazard regression models were applied for MFS analysis on the expression levels of individual genes (Harrington and Fleming, 1982; Therneau and Grambsch, 2000). The genes with hazard ration (HR) above 1 are considered as risk genes and genes with HR$<$ 1 are defined as protective genes.

For the multiple testing problem Bonferroni method was used when patient data were subsetted resulting in q-values and FDR method was applied when correcting gene significance in a tested gene list.

2.3 Simulations

Synthetic expression data were generated for two extensive simulation studies. Each study comprised five simulation types (Table 2.4), which differed in the topology designs for pathway deregulation (see section 2.3.3) and in a choice of the parameter whose ranges were investigated (see section 2.3.2). The two distinct simulation studies differed in their gene ID annotation. Whereas in the first study real gene HGNC symbols were used, in the second study these symbols were replaced by unique synthetic IDs for all nodes of all pathways in order to prevent overlapping of the pathways.

Expression data were drawn from a multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is p-dimensional mean vector and $\boldsymbol{\Sigma}$ is $(p \times p)$-dimensional covariance matrix. Balanced sample size was always generated consisting of 10 control and 10 treatment samples. As treatment samples I refer to the group in which changes in the mean vector were introduced. By this, expression levels of the genes, which reflect fold-changes between the two groups, were controlled. The genes that had an increased expression above 0 are called affected. In the covariance matrix a correlation of 0.8 between genes in the same pathway was introduced and 0.05 otherwise. Variance in the matrix was set to 2. For the generation of both treatment and control samples, the same covariance matrix was used.

2.3.1 Studies with different pathway input

The differences between simulation study 1 and 2 originate from their different pathway data inputs.

Within the study 1, original pathways with overlapping genes were used. When generating expression data for the study 1, the dimension of mean vector was set to $p = 3173$. This number represents the total number of unique gene symbols in parsed KEGG pathway graphs. As the genes in this study can belong to multiple pathways the covariance matrix was not positive-definite and its close positive-definite approximation was computed using the *sfsmisc* R-package.

In the study 2 non-overlapping pathways with unique gene IDs were generated; however, with the same pathway topology as the original pathways. To simulate expression data in the second study the mean vector dimension was $p = 7697$. This number represents the total number of nodes in the KEGG pathway graphs. The covariance matrix of was constructed in a similar manner as in the first study, but as pathways do not overlap the matrix was positive-definite by default.

2.3.2 Variable parameters

In each study five simulation types were performed. Within a single simulation type one of the four variable parameters was explored: mean vector (*mean*), pathway size (*size*), number of pathway (N) and detection call (DC). Several levels of a variable parameter were investigated, while the remaining three parameters were fixed at a certain level. This resulted in 17 distinct configurations of parameters (Table 2.4). Each parameter configuration was examined in 1000 simulation runs, meaning that 1000 expression matrices were generated for the given configuration.

Conf.	**Type**	**Topology design**	**Variable parameter**	**Levels of the parameters**			
				mean	*size*	N	DC
1	1	Community	*mean*	±1	all	12	50%
2				±2			
3				±6			
4	2		*size*	±2	small	12	50%
5					median		
6					big		
7	3		N	±2	all	12	50%
8						23	
9						70	
10	4	Betweenness	DC	±2	all	12	10%
11							30%
12							50%
13							70%
14	5	Neighbourhood		±2	all	12	10%
15							30%
16							50%
17							70%

Table 2.4. *Five simulation types with 17 different parameter configurations, comprising different combinations of a topology design and a variable parameter. Table adapted from Bayerlová et al. (2015a).*

Mean vector

To investigate the influences of mean vector changes I choose three levels $mean = \{\pm 1, \pm 2, \pm 6\}$. These expression changes were introduced for affected genes into the mean vector of the treatment group. The mean vector of the control group always remained 0. After the differential analysis, the different magnitude of *mean* change is reflected in logarithm of fold-change (logFC) of a gene in between two conditions. The direction of expression change in the *mean* was positive (+) for one half and negative (−) for the second half of the affected genes. When ranges of other parameters were investigated, the default of expression change was set to $mean = \pm 2$.

Pathway size

The second parameter was size of the deregulated pathways and its three levels comprised $size = \{\text{small, medium, big}\}$. All pathways in the parsed KEGG pathway data input were stratified according to their number of genes into three size groups. Small pathways contain from 5 to 26 genes (minimum to 25% quantile), medium pathways where considered when having from 27 to 85 genes and big pathways consist of 86 up to 380 genes (75% quantile to maximum). Pathways of all sizes were taken into account, when *size* parameter was set as fixed.

Number of pathways

Next, I was interested in studying the effect of different proportions of the whole pathway database being deregulated. The number of deregulated pathways $N = \{12, 23, 70\}$ represent approximately 10%, 20% and 60% of all pathways in the KEGG database, respectively. Ten percent of all pathways ($N = 12$) were assigned as deregulated when another parameter was variable.

Detection call

The so-called detection call (DC) (Tripathi and Emmert-Streib, 2012) defines the percentage of affected genes in a deregulated pathway. Four levels of DC were investigated $DC = \{10\%, 30\%, 50\%, 70\%\}$ to explore how many

genes in a pathway have to be assigned as affected in order to detect this pathway as significant. Half of the genes in a deregulated pathway were affected ($DC = 50\%$) when this parameter was not variable in a given simulation type.

2.3.3 Topology designs for pathway deregulation

To deregulate a pathway within a stimulation, some of its genes needed to be affected – to have altered expression in treatment samples compared to controls. I examined thee approaches how to reflect pathway topology to allocate affected gene in a deregulated pathway: community, betweenness and neighbourhood approach (Figure 2.2, for general definitions of these measures see Introduction section 1.6.2 *Network analysis*).

The communities were detected by maximizing modularity measure of the pathway graph and for each pathway I searched for a community which represented from 45% to 55% DC. However, for some pathways several communities had to be joined or a too big community had to be cut to achieve the appropriate DC. To select affected genes in the betweenness deregulation design I considered the top highest scored betweenness nodes. The required number of the top scored nodes for the given DC was assigned as affected. In the neighborhood deregulation approach one node of a pathway was chosen and all nodes within certain distance created the neighborhood. In the search for the neighborhoods representing different ranges of DC parameter, the neighborhoods of several orders for each pathway node were calculated and the one best fitting the required DC level ($\pm 5\%$) was chosen.

2.3.4 Single simulation run

Within a given parameter configuration for a run, a certain number of pathways (depending on the N parameter – when fixed $N = 12$ pathways) were randomly chosen to be deregulated by assigning affected genes. Then new expression data for this particular run were drawn from the multivariate normal distribution. The expression matrix was directly supplied to the CePa ORA and CePa GSA algorithms. For the rest of the methods linear models were fitted first, in order to identify DEGs. Those were further supplied to the enrichment methods including logFC, p-values or FDRs, according to the specific requirements of

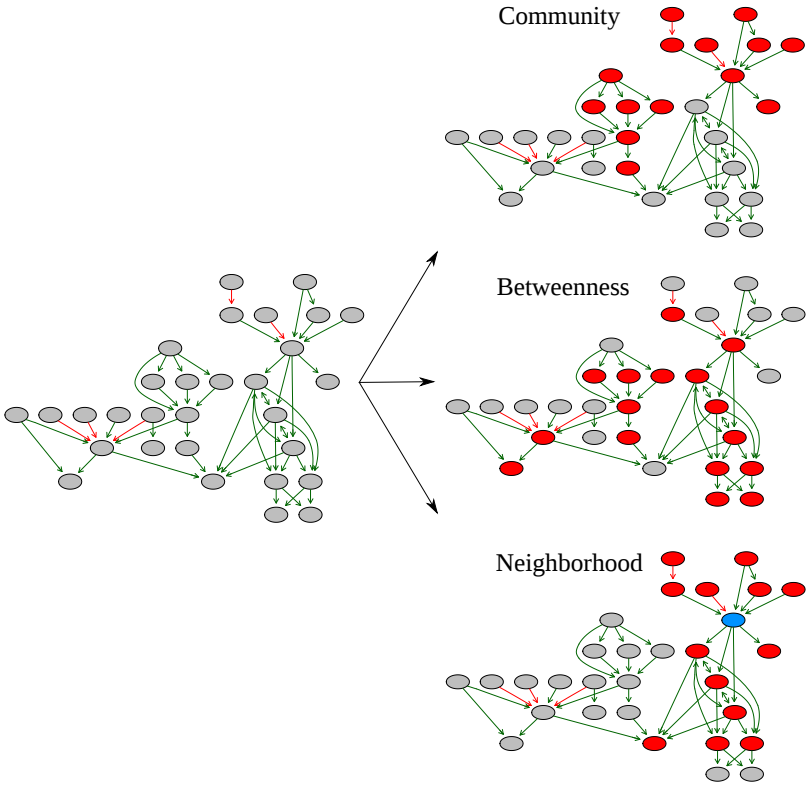

Figure 2.2. *Example of a pathway with 30 genes. In order to deregulate this pathway on detection call level DC = 50% (±5%) I needed to assign 14-16 affected gene to this pathway and allocate them on the pathway graph according to three topology designs for pathway deregulation: In the community design two gene communities were selected, together comprising 14 affected nodes (depicted in red). Using betweenness approach, 16 top scored nodes were affected. Neighborhood of order 2 of the blue node was selected resulting in 16 affected nodes. The pathway edges represent activation (green) and inhibition (red). Figure adapted from Bayerlová et al. (2015a).*

each method. In each simulation run all 7 methods were evaluated on the basis of the same expression data and pathway data inputs.

2.3.5 Evaluation

The simulations were evaluated in the terms of sensitivity, specificity and accuracy (see Table 2.5). The sensitivity describes the proportion of true-positive pathways detected out of all pathways assigned as deregulated. The specificity describes the proportion of true-negative pathways out of all pathways that were not deregulated (i.e. without assigned affected genes). The accuracy

is given as the proposition of true-positive and true-negative pathways detected by a method out of all tested pathways.

		Reality - Is pathway affected?		
		Yes	No	
Test - Is pathway enriched? $FDR < 0.05$	Yes	TP	FP	
	No	FN	TN	
		Sensitivity $\frac{TP}{TP+FN}$	Specificity $\frac{TN}{FP+TN}$	Accuracy $\frac{TP+TN}{TP+FP+FN+TN}$

Table 2.5. *Measures for performance evaluation: definition of sensitivity, specificity and accuracy.*

Results

This chapter presents first the parsed prior knowledge on signaling pathways suitable for further integration into bioinformatic analysis. Subsequently, I compare performance of different enrichment analysis approaches, describe the features of newly constructed Wnt signaling networks, present results of network integration with newly sequenced breast cancer cell line data, and bring these results into the clinically relevant context of metastatic breast cancer expression profiles.

3.1 Parsed pathways

Prior pathway knowledge was required for creating Wnt signaling models as well as comparing enrichment methods. All pathway information utilized in this work was parsed from Biological Pathways Exchange (BioPAX) exports of public pathway databases. Resulting interaction graphs of the pathways had to be further edited (see 2.2.3 *Parsing pathway knowledge from databases*) in order to associate a single graph node to a single gene ID as well as to preserve signal propagation continuity. The final directed graphs modeled the pathways as cascades of activating and inhibiting processes between the genes.

In order to assemble comprehensive list of Wnt signaling and Wnt-related pathway five databases were scanned yielding a collection of 26 interaction graphs (Appendix Table 1). These were further merged into signaling networks described in the 3.3 *Wnt networks* section.

For enrichment methods comparison 116 signaling as well as disease pathways from the KEGG database were selected omitting metabolic pathways. Final interaction graphs contained 7697 nodes in total, comprising 3173 unique genes. This reflects the extent of pathway overlap as one gene was present on

average in 2.4 pathways. The individual pathways comprised from 5 up to 380 genes with a median pathway size of 55.5 genes (Figure 3.1). Despite the effort to ensure connectivity of a graph when non-gene molecules were filtered out, a single pathway did not always consist of a single connected component. The median number of connected components in an interaction graph was 2. The KEGG interaction graphs were further customized into appropriate pathway input for the individual enrichment analysis methods.

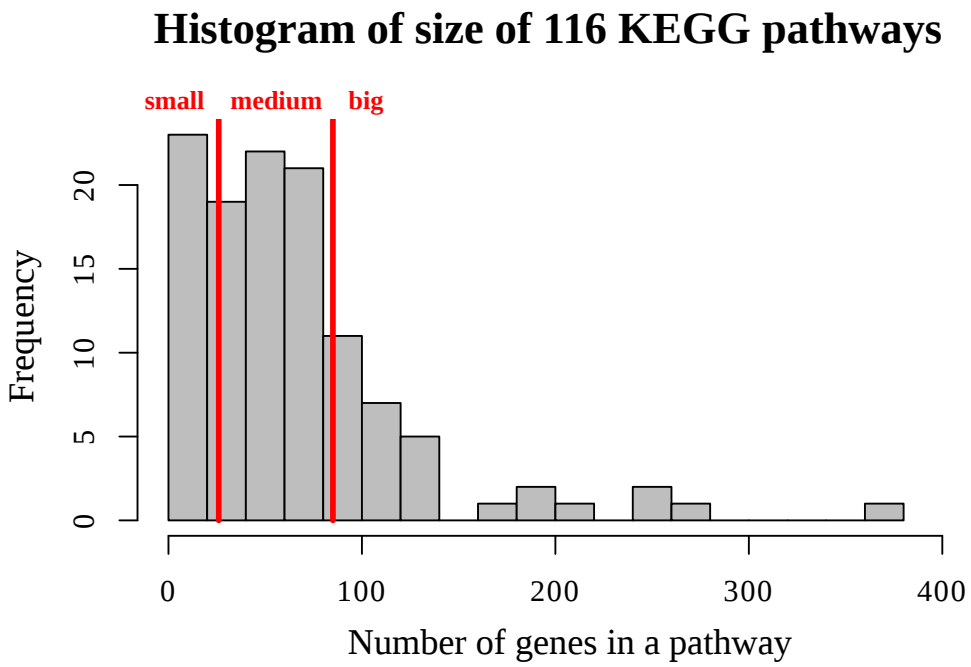

Figure 3.1. *Parsed KEGG database consist of 116 pathways, with the smallest pathway of 5 genes and the biggest of 380 genes. Red lines depict 25% and 75% quantiles which served as ranges for pathway size parameter in the simulation studies.*

3.2 Comparison of enrichment methods

Seven enrichment methods (Table 2.3) representing gene-set (GS) as well as pathway topology-based (PT-based) analysis approach were compared in several simulation scenarios on synthetic expression data generated under different

parameter settings. To ensure comparability of the results the same KEGG database export was used for all methods to compile gene sets and various pathway graph representations. Further, the algorithms were applied and evaluated on real microarray data comprising 24 disease datasets.

3.2.1 Simulation studies

In the simulation studies the methods were evaluated in term of sensitivity and specificity. Within each study 17 parameter configurations (Table 2.4) were considered and for each configuration the median sensitivity and specificity of 1000 simulation runs were computed. The parameters and their settings are described in detail in 2.3 *Simulations* section.

3.2.1.1 Study 1: with original pathways

Figure 3.2 depicts the results of simulation study with the original overlapping KEGG pathways.

With weak changes of the mean vector ($mean = \pm 1$) sensitivity of GS methods based on ranking was the best: 0.67 for Wilcoxon rank sum (WRS) and 0.58 for Kolmogorov-Smirnov (KS), followed by PathNet with a sensitivity of 0.5. All methods based on over-representation analysis (ORA) (both GS and PT-based) had 0 sensitivity; however, that was reflected in a specificity of 1. With a higher mean change, i.e. $mean = \{\pm 2, \pm 6\}$, all methods were comparably sensitive (between 0.92 and 1), whereas the best specificity scores (0.51 and 0.49) were reached by WRS. In the case of small pathways being deregulated, most of the methods were less sensitive than in detecting the bigger pathways. However, CePa GSA (self-contained) and WRS performed best with a sensitivity of 0.92 and 0.83, respectively. Specificity decreased for the medium and big pathways. Especially CePa GSA was very unspecific on the big pathways (0.22) in comparison to the other methods, whose specificity ranged from 0.46 to 0.56. When more than half of the database pathways were deregulated ($N = 70$) both sensitivity and specificity decreased compared to $N = \{12, 23\}$. For $N = 70$ CePa GSA specificity was only 0.087 whereas the others ranged between 0.3 and 0.5. At low level of detection call ($DC = 10\%$)

coupled with betweenness topology design the PT-based methods and Fisher's exact (FE) test were more sensitive ($0.58 - 0.92$) than ranking GS methods ($0.33 - 0.58$). However, on 30% and higher detection call (DC) levels all methods performed comparably well in the term of sensitivity ($0.83 - 1$). At the same low level of detection call ($DC = 10\%$) coupled with neighborhood topology design PathNet had the highest sensitivity of 0.58, followed by other PT-based methods and FE ($0.42 - 0.5$). In comparison to the corresponding simulation type with betweenness design, in this setting only CePa ORA and CePa GSE reached a sensitivity over 0.8 for $DC = 30\%$.

The sensitivity was fairly high across the configurations and rather comparable for both GS and PT-based methods. However, the specificity over all parameter configurations was very low for all methods. Thus, I performed the same five simulation types on the non-overlapping pathways in the Study 2.

3.2.1.2 Study 2: with non-overlapping pathways

Figure 3.3 depicts the results of simulation study with the non-overlapping KEGG pathways, whose topology was preserved but the nodes were assigned with unique synthetic IDs. This resulted in markedly better specificity, whereas the sensitivity results tell rather similar story as in the Study 1.

On the low level of mean parameter ($mean = \pm 1$) again the GS methods based on ranking were the most sensitive (0.5). In detecting small pathways the best sensitivity (0.75) was reached by WRS and PathNet. Whereas for the big pathways sensitivity of WRS was 1 and of PathNet 0.92. The sensitivity of all methods except CePa GSA decreased when many pathways were deregulated ($N = 70$), even to a bigger extent than in Study 1. Interestingly, FE and KS specificity was only 0.17 for $N = 70$, while most of the other methods reached a specificity of 1. When pathways did not overlap the DC level of 50% in the betweenness design was needed to achieve sensitivity over 0.8, whereas for the original pathway study all methods reached 0.8 sensitivity at a $DC = 30\%$ level. For the low DC ($DC = 10\%$) PathNet's sensitivity was the best (0.58). Simulation type of DC coupled with neighborhood topology design had a similar behavior pattern as the betweenness setting; only for the both CePa methods sensitivity on $DC = 10\%$ level decreased to 0.

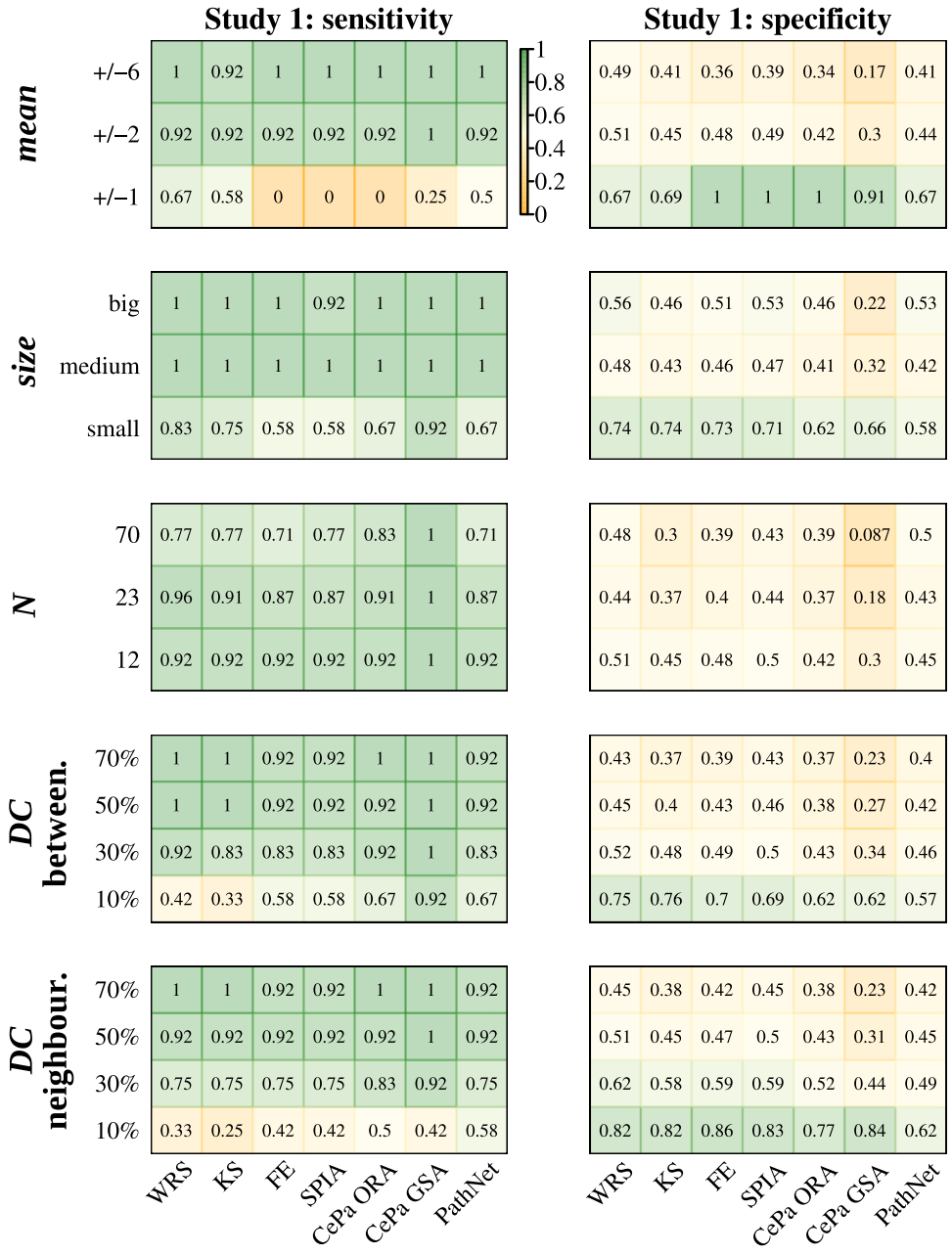

Figure 3.2. *Simulation study 1 using original pathways with overlapping genes: Sensitivity and specificity scores of seven methods under 17 parameter configurations. Each cell summarizes a median value of 1000 runs. The same color code key applies for all simulation types. Figure adapted from Bayerlová et al. (2015a).*

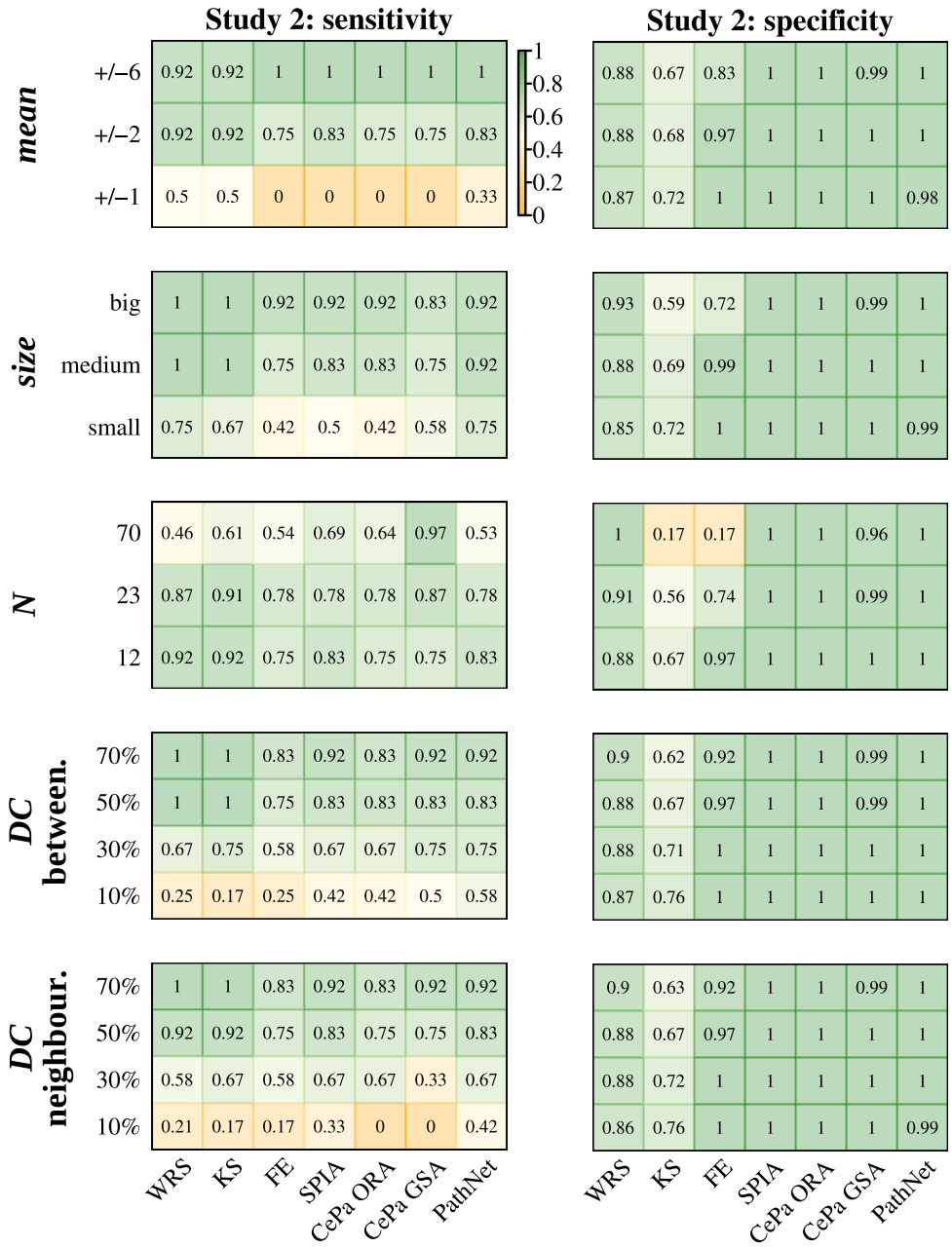

Figure 3.3. *Simulation study 2 using non-overlapping pathways with unique gene IDs: Sensitivity and specificity scores of seven methods under 17 parameter configurations. Each cell summarizes a median value of 1000 runs. Figure adapted from Bayerlová et al. (2015a).*

3.2.1.3 Overall performance in simulations

Average sensitivity, specificity, and accuracy were calculated across 17 configurations for each method in both studies (Figure 3.4). In the first simulation study with overlapping pathways the average specificity was very low for all methods, achieving maximum of 0.55 for WRS and SPIA. This was also reflected in the best accuracy of these two methods: 0.6 for WRS and 0.59 for SPIA. The rest of the methods achieved comparable accuracy, with levels between 0.53 and 0.57, with the exception of CePa GSA, which showed an average accuracy score of 0.46. In the second simulation study with non-overlapping pathways, both overall specificity and accuracy increased. For PT-based methods the average specificity was 1, whereas there were prominent differences within GS methods; average specificity for both WRS and FE was 0.89, for KS it was only 0.65. The sensitivity in Study 2 varied moderately among the methods from 0.63 up to 0.77. The best overall accuracy in the non-overlapping pathways setting was achieved by the four PT-based methods ($0.95 - 0.96$), followed by FE with 0.88 and WRS with 0.86 scores.

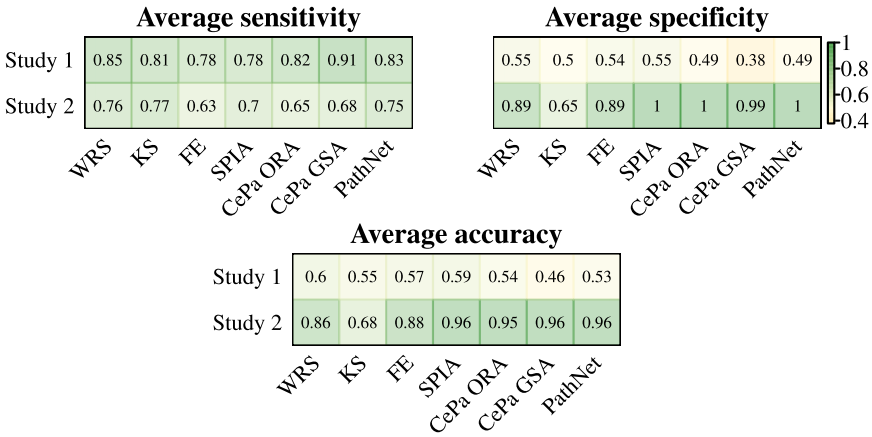

Figure 3.4. *Overall sensitivity, specificity and accuracy in the two simulation studies: Mean of three measures for each method over 17 configurations.*

3.2.2 Performance in benchmark data

Besides simulated expression data, 24 microarray datasets were used as a benchmark to compare performance of seven enrichment methods. Each disease

dataset had its defined target pathway from the KEGG database (Table 2.1). As the target pathway is only one of several potential true positive pathways for a given disease, I evaluated neither sensitivity nor accuracy. Instead, the p-values of the target pathways in 24 datasets and their ranks in the whole KEGG database were inspected, expecting the target pathways to be ranked close to the top. The lowest p-values of the target pathways were detected by CePa GSA, followed by WRS (Figure 3.5A). In the ranking of the target pathways the PathNet method performed best (Figure 3.5B). Figure 3.5C shows the proportion of the significant to non-significant pathways detected within the 24 datasets. The CePa GSA identified on average 68% of all database pathways as significantly enriched, whereas for the rest of the methods it was between 1.7% and 7.4% of the significant pathways.

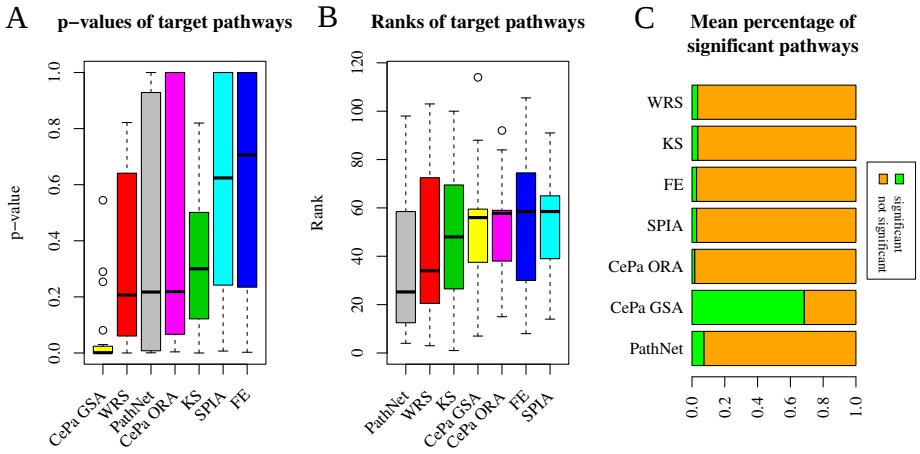

Figure 3.5. *Comparison of 7 methods on the benchmark data: Distributions of p-values and ranks of the target pathways in 24 datasets. Methods are ordered according to the median p-value (A) and rank (B) from the best to the worst – lower values better performance. (C) Average percentage of the 116 pathways detected as significant and not significant by each method.*

3.3 Wnt networks

To answer the second major question within my thesis – *Which module of the Wnt signaling network is active in aggressive breast cancer?* – first, the network representation of Wnt signaling was needed.

Pathway databases provide information on Wnt signaling cascades to differing extents. Scanning the databases for the pathways linked to Wnt signaling resulted in 26 parsed pathways. These were sorted into four conceptual groups (Appendix Table 1) reflecting different Wnt signaling pathway. The first two groups were defined straightforwardly as (1) *canonical* and (2) *non-canonical Wnt signaling*. However, several pathways did not fall into this classification. After inspecting these pathways more closely[(1)], two additional groups were created: For the pathways which act upstream of the Wnt cascade (3) the *regulation of Wnt signaling* was defined. And for the pathways which reflect the *off-state* of the Wnt signaling, when β-catenin is targeted by the destruction complex, (4) the *inhibition of canonical Wnt signaling* group was established.

The ensemble models were created by merging the interaction graphs within each group (see Figure 2.1C2). This merging yielded four signaling networks: canonical Wnt signaling, non-canonical Wnt signaling, inhibition of canonical Wnt signaling, and regulation of Wnt signaling (Figure 3.6).

The networks were inspected for their general features (see 3.3.1 *Network properties*) and for the structural relationships of well-established Wnt pathway players (see 3.3.2 *Clustered Wnt subnetwork*).

3.3.1 Network properties

The first step to describe the networks was to assess their size. Surprisingly, although the focus of scientific publications seems to be imbalanced in favor of canonical Wnt signaling, the size of the non-canonical network (489 nodes, 7869 edges) exceeded that of the canonical network (304 nodes, 2686 edges). The regulation of Wnt signaling network comprised 173 nodes with 589 edges, whereas the inhibition of canonical signaling network had just 83 nodes but were richly interconnected by 675 edges. Due to the considerable size of the networks, systematic visual exploration was not suitable. To further investigate the features of the four networks I identified their *key nodes* – the most central genes defined by degree and betweenness measures (see the nodes depicted in black in Figure 3.6).

(1) Consulting with Dr. med. Florian Klemm and Dr. med. Annalen Bleckmann, Department of Hematology and Medical Oncology, University Medical Center Göttingen.

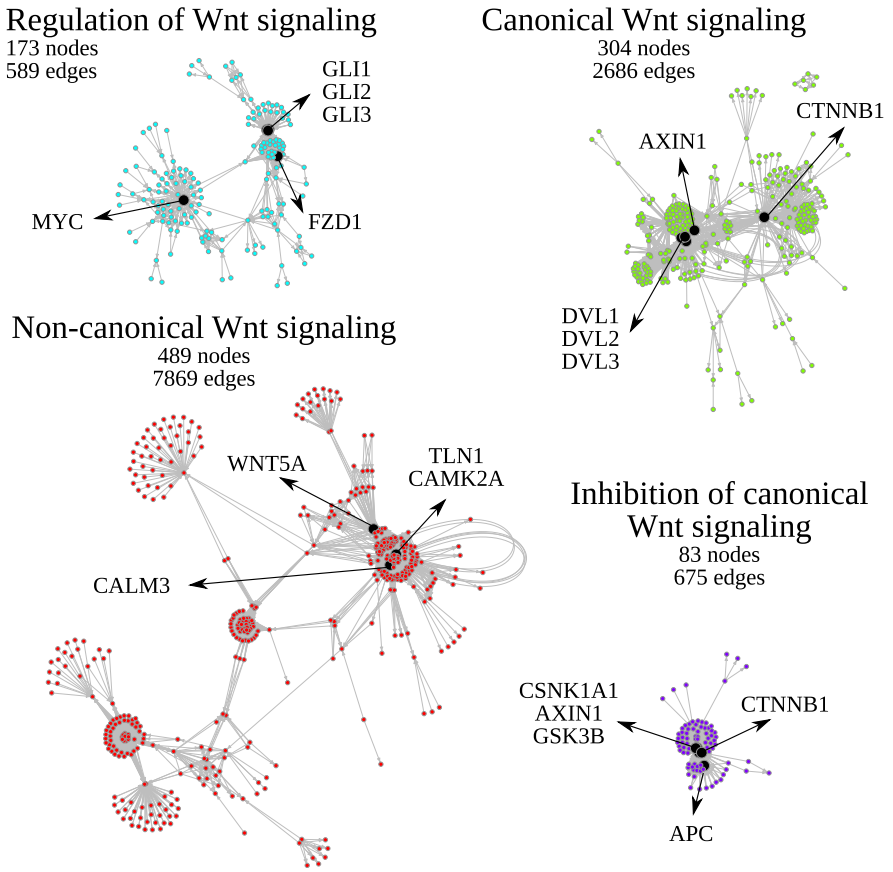

Figure 3.6. *Four Wnt networks representing different Wnt signaling pathways. The size of each network is defined by the numbers of nodes and edges. A maximum number of five top key nodes are visualized in black for each network. Figure adapted from Bayerlová et al. (2015b).*

3.3.2 Clustered Wnt subnetwork

To ascertain the biological relevance of the newly constructed networks, I narrowed the focus from the large-scale networks into well-known Wnt pathway genes preset therein. I collected a list of 121 established Wnt pathway players reviewed in the literature (Moon et al., 2004; Katoh and Katoh, 2007; Amerongen and Nusse, 2009; Holland et al., 2013) and overlaid them onto nodes of the networks. The subnetwork was extracted comprising 112 mapped genes interconnected by 931 edges from all four networks. Subsequently, the subnetwork of reviewed Wnt genes was subjected to structural analysis. Partitioning the sub-

network by increasing the modularity revealed five well connected topological components, so-called communities (Figure 3.7).

The first community (in purple) comprised mostly Wnt and secreted frizzled-related proteins representing a cluster of ligands. Two side communities constituted β-catenin-independent signaling: Wherein the smaller (yellow) community consists of players from the Planar cell polarity (PCP) pathway, the second (green) cluster included *WNT5A* and *WNT11* and further downstream components of the Wnt/Ca^{2+} pathway. The central (blue) community was more heterogeneous, as it contained both canonical (e.g. *WNT3A*, *LRP5*, *LRP6*, *AXIN1*, *APC*, *CTNNB1*) and non-canonical (e.g. *VANGL1*, *RHOA*, *DAAM1*) components. The central position of Dishevelled (Dvl) isoforms nodes in this community is consistent with their roles in transducing both canonical and non-canonical Wnt signals. Finally, the bottom red community grouped the transcriptional effectors and target genes.

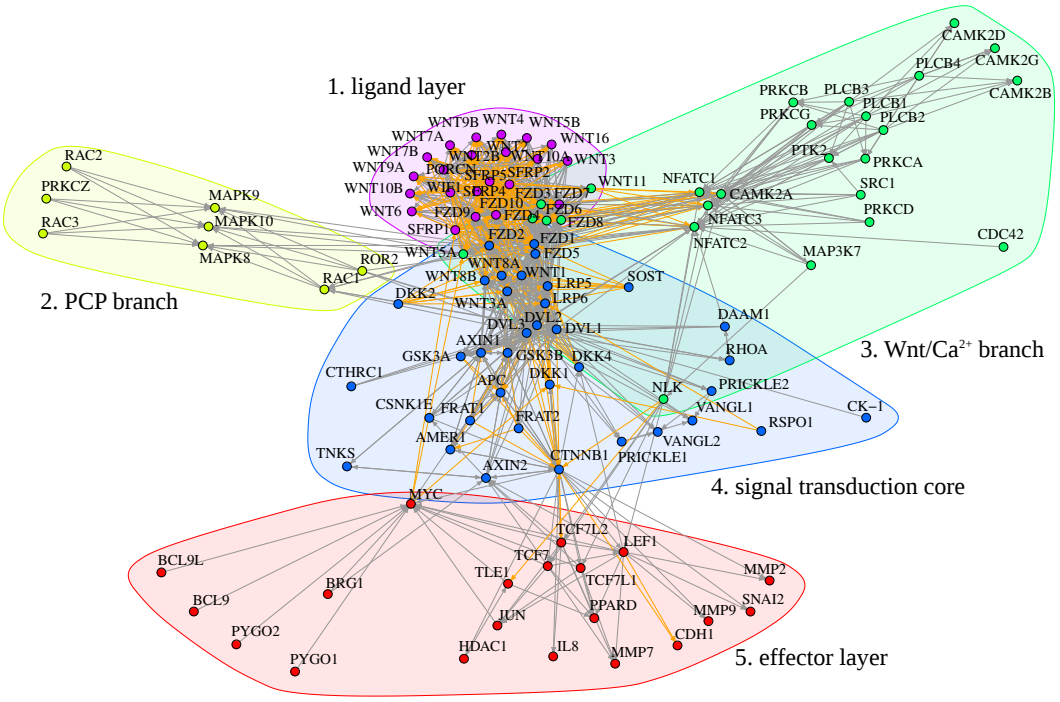

Figure 3.7. *Subnetwork of reviewed Wnt genes: The five communities marked in different colors were detected by optimizing the modularity measure. The directed edges represent activating (gray) and inhibiting (orange) interactions. Figure adapted from Bayerlová et al. (2015b).*

3.4 Sequenced cell line and perturbation targets

RNA-Seq data were generated in order to explore the downstream effects of different interventions in the estrogen receptor (ER) positive MCF-7 breast cancer cell line. The phenotype of the MCF-7 cells without any intervention experiments is considered to correspond to luminal A breast cancer subtype with favorable clinical prognosis. The working hypothesis within this context is that activation of non-canonical Wnt pathway stimulates cell proliferation and migration. Wnt5a ligand and Ror2 receptor were chosen as non-canonical Wnt pathway members for the perturbation experiments in order to enhance invasiveness of MCF-7 cell line and subsequently to identify expression-responsive target genes. Four conditions of MCF-7 cell line were generated:

1. positive control (ctl) with empty pcDNA vector
2. control stimulated with Wnt5a (ctl + Wnt5a)
3. stable over-expression of Ror2 receptor (Ror2)
4. combination of both perturbations (Ror2 + Wnt5a)

The MCF-7 cells stimulated with Wnt5a showed enhanced cell invasiveness compared to the control. The Ror2 over-expression also resulted in an increase of tumor cell invasion and this could be further stimulated by treatment with Wnt5a in parallel (Figure 3.8).

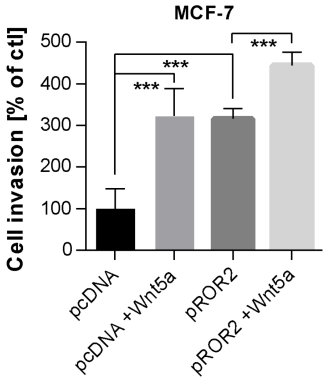

Figure 3.8. *In-vitro microinvasion assay of MCF-7 cells in four conditions: The number of invasive cells is related to the control empty vector condition (pcDNA). Statistical significance* $p < 0.001$ *is indicated as ***. This figure was provided by Dr. med. Annalen Bleckmann and Dr. Kerstin Menck.*

To quantify changes in gene expression underlying this phenotypic transformation each MCF-7 condition was deep sequenced in three replicates. Based on RNA-Seq data the gene-level abundance were estimated and the total library size of sequenced samples ranged from 35 to 55 million (Table 3.1). In the differential analysis transcriptomic profiles of the perturbed cell lines were compared to the control samples or other perturbed conditions of interest.

Table 3.2 summarizes the genes differentially expressed between the compared conditions. The two comparisons testing for the effect of Wnt5a stimulation – with and without presence of the over-expressed Ror2 – yielded rather low numbers of significantly differentially expressed genes (DEGs) (Figure 3.9, Appendix Table 2). The only significant differential gene detected in the both comparison was *MUC5AC*, which indicated that Wnt5a stimulation had only moderate effect on the transcriptomic level changes of the MCF-7 cell line. Out of five performed comparisons, the three comparisons that tested for Ror2 over-expression impact (ctl vs. Ror2, ctl vs. Ror2 + Wnt5a, ctl + Wnt5a vs. Ror2 + Wnt5a) demonstrated stronger effects resulting in 2860, 3729 and 3022 DEGs, respectively. To identify stable targets of Ror2 over-expression I determined the overlap of these three gene lists in a venn analysis that resulted in 2068 common targets (Figure 3.10).

Condition	**Rep.**	**Library (mio)**
ctl	1	50
ctl	2	53
ctl	3	47
ctl + Wnt5a	1	35
ctl + Wnt5a	2	55
ctl + Wnt5a	3	49
Ror2	1	48
Ror2	2	36
Ror2	3	53
Ror2 + Wnt5a	1	49
Ror2 + Wnt5a	2	51
Ror2 + Wnt5a	3	35

Table 3.1. *Deep sequenced sample replicates of perturbation experiments on MCF-7 with the total number of the mapped reads in million (mio).*

Comparison of conditions			N. of DEGs
ctl	vs.	ctl + Wnt5a	7
ctl	vs.	Ror2	2860
ctl	vs.	Ror2 + Wnt5a	3729
Ror2	vs.	Ror2 + Wnt5a	11
ctl + Wnt5a	vs.	Ror2 + Wnt5a	3022

Table 3.2. *By comparing transcritomic profiles of the MCF-7 cell lines with different perturbation, significantly differentially expressed genes were identified between the corresponding conditions (Number of DEGs).*

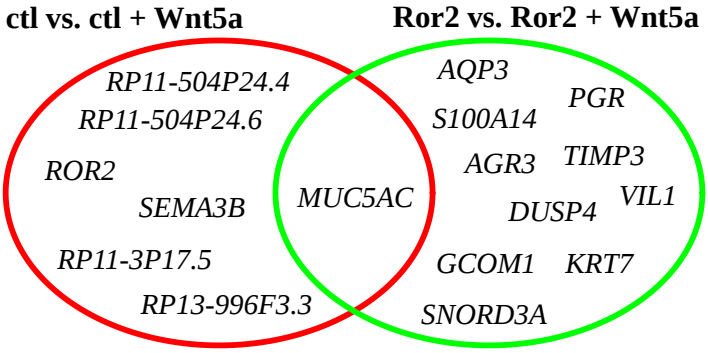

Figure 3.9. *Targets of Wnt5a stimulation in the MCF-7 cells with and without Ror2 receptor over-expression.*

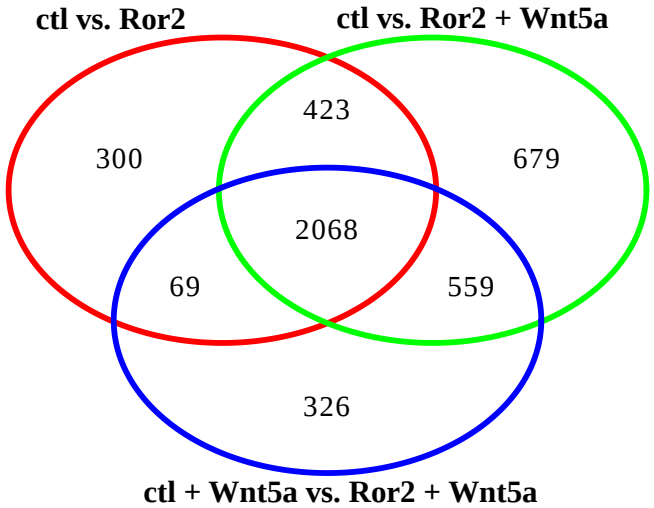

Figure 3.10. *Common Ror2 targets: Venn diagram depicting overlap of three lists of differential genes responsive to Ror2 over-expression in MCF-7.*

3.4.1 Enrichment of Ror2 targets in Wnt gene sets

The enrichment analysis was performed to determine which Wnt signaling cascades were activated within the Ror2 over-expressing system. The list of 2068 common targets represents the threshold setting, therefore the ORA approach was utilized. The four constructed Wnt models were used as prior Wnt pathway knowledge in the form of gene sets. To further explore the contribution of the up-regulated and down-regulated genes to the enrichment, the targets were split based on positive and negative fold-changes into three lists: *all*, *up* and *down*. These sets of targets were used to determine the enrichment patterns of the Wnt pathway gene sets (Table 3.3). Two Wnt pathways were detected as over-represented in the list of all targets: Non-canonical and regulation of Wnt signaling. Whereas the non-canonical Wnt gene set was significant for all as well as for up-regulated targets, the canonical Wnt gene set was not significant for any target list.

Wnt model as a gene set	**2068 targets**		
	All	**Up**	**Down**
Canonical Wnt signaling	0.35	0.3	0.54
Non-canonical Wnt signaling	**0.02**	**0.002**	0.65
Inhibition of canonical Wnt signaling	0.98	0.91	0.97
Regulation of Wnt signaling	**0.01**	0.09	**0.04**

Table 3.3. *Over-representation analysis of the common targets: FE test was performed to test for over-representation of 2068 targets and their up-regulated and down-regulated subsets in the distinct Wnt gene sets. Significant p-values ($p < 0.05$) are depicted in bold.*

3.5 Integration of targets with networks

The list of 2068 targets was further used to integrate the results of RNA-Seq experiments into the framework of networks. Within the Wnt knowledge context the enrichment analysis results indicated activation of the non-canonical Wnt model in the gene expression data. Therefore, the non-canonical Wnt network was chosen for the subsequent network integration analysis to identify a module activated by the Ror2 over-expression targets (see 3.5.1 *Non-canonical Wnt module*). Moreover, to explore the 2068 targets besides the Wnt signaling context, I integrated them with public protein-protein interaction (PPI) network from the BioGRID database (see 3.5.2 *PPI network and hubs*).

3.5.1 Non-canonical Wnt module

As the non-canonical Wnt model exhibited significant enrichment of the common targets when utilized as the gene set, I further exploited its network structure to identify the expression-responsive module. First, 2068 genes were mapped onto the nodes of non-canonical network, which resulted in 66 network nodes induced by the targets. To link these 66 nodes within the network structure, the Steiner tree algorithm was employed, which introduced 18 connecting nodes that do not embody differential targets (Appendix Table 3). Subsequently, the induced subnetwork of 84 nodes was extracted including all original edges (Figure 3.11, Appendix Table 4). This subnetwork represents the module of non-canonical Wnt pathway regulated by over-expressed Ror2 receptor (hereinafter referred to as *Wnt module*). The module revealed several important Wnt pathway members: *FZD5* and *WNT11* as up-regulated and *FZD4* and *DVL1* as down-regulated in response to Ror2 over-expression, and *WNT5A*, *DVL2* and *CD36* as Steiner nodes interconnecting the differential targets.

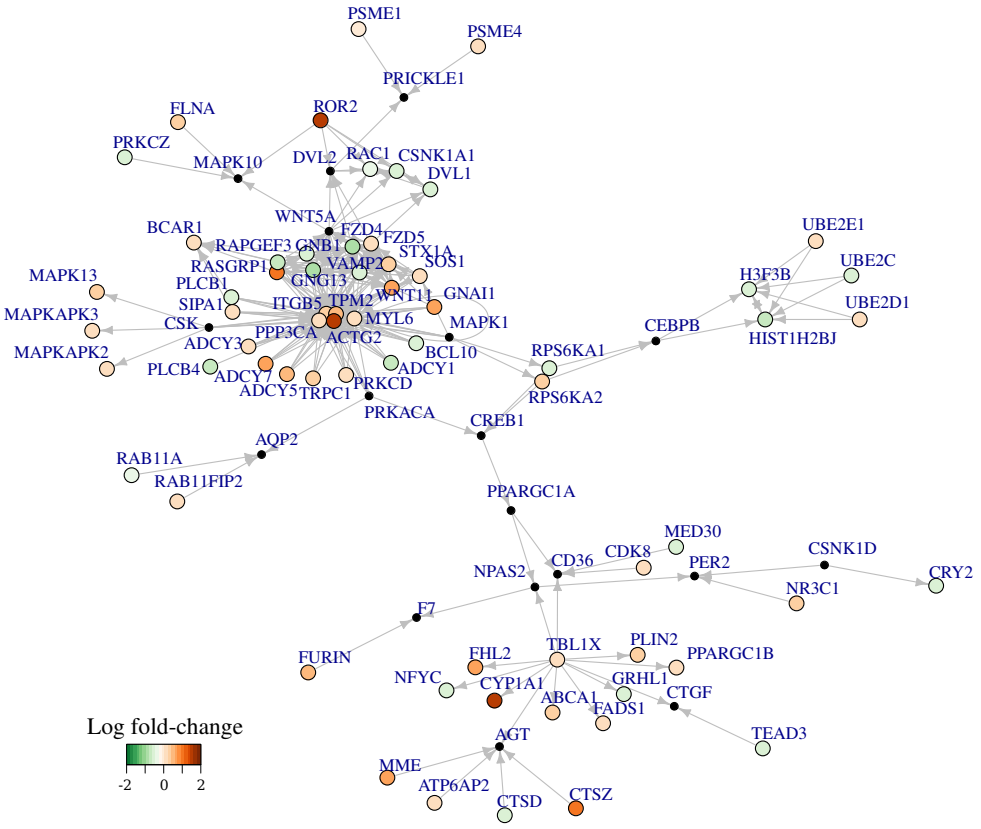

Figure 3.11. *Non-canonical Wnt subnetwork representing differentially regulated module of Wnt pathway. The color coding corresponds to the fold-changes of the targets: green represents down-regulation and brown up-regulation of the genes in Ror2 over-expressing cells compared to the control MCF-7 cells. The smaller black nodes were introduced by the Steiner tree analysis. The directed edges represent controlling interactions.*

3.5.2 PPI network and hubs

To investigate which of the targets have important functional relationships within cellular network of proteins, I integrated them into the interactome represented as undirected network of PPIs. From 2068 genes 1710 were mapped onto the PPI network and only edges connecting induced nodes were preserved. The induced graph contained one major connected component, 13 small components

of three or less nodes, and 767 genes were interaction orphans. I focused on the biggest component comprising 920 targets interconnected by 2199 edges. To explore relevant players within this PPI graph component, I used two centrality measures to identify hub genes (described in 2.2.5 *Network-based analyses*). 41 hubs were detected that represented nodes with central topological position – meaning they had many interaction partners from the target genes as well as they facilitated the transfer between the targets (Figure 3.12). Table 3.4 summarizes the 41 hubs which include, for instance estrogen and androgen receptors (*ESR1*, *AR*), (proto-)oncogenes (*MYC*, *BMI1*, *FYN*), as well as four genes overlapping with Wnt module members: *UBE2D1*, *FLNA*, *RAC1*, and *FHL2*.

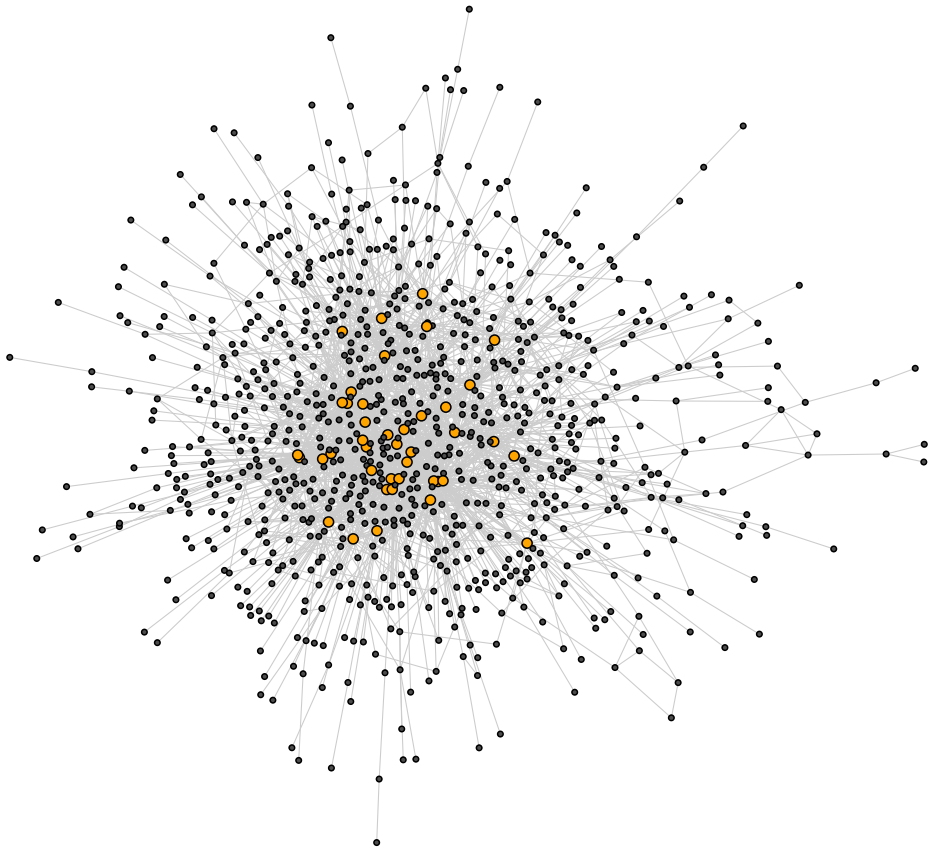

Figure 3.12. *Hub targets in PPI network: Connected component of 920 targets with 2199 PPIs with central hub nodes highlighted in orange.*

	Hub gene	Description
1	*FN1*	fibronectin 1
2	*ESR1*	estrogen receptor 1
3	*COPS5*	COP9 signalosome subunit 5
4	*HDAC1*	histone deacetylase 1
5	*MYC*	v-myc avian myelocytomatosis viral oncogene homolog
6	*BMI1*	BMI1 proto-oncogene, polycomb ring finger
7	*SUMO3*	small ubiquitin-like modifier 3
8	*RPA2*	replication protein A2, 32kDa
9	*YWHAG*	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, gamma
10	*EZH2*	enhancer of zeste homolog 2 (Drosophila)
11	*PAN2*	PAN2 poly(A) specific ribonuclease subunit homolog (S. cerevisiae)
12	*UBE2I*	ubiquitin-conjugating enzyme E2I
13	*SH3KBP1*	SH3-domain kinase binding protein 1
14	*SOX2*	SRY (sex determining region Y)-box 2
15	*BARD1*	BRCA1 associated RING domain 1
16	*HDAC5*	histone deacetylase 5
17	*ATXN1*	ataxin 1
18	*SRRM2*	serine/arginine repetitive matrix 2
19	*ARRB1*	arrestin, beta 1
20	*KDM1A*	lysine (K)-specific demethylase 1A
21	*LGR4*	leucine-rich repeat containing G protein-coupled receptor 4
22	*SFPQ*	splicing factor proline/glutamine-rich
23	*CD81*	CD81 molecule
24	*AR*	androgen receptor
25	*PRKDC*	protein kinase, DNA-activated, catalytic polypeptide
26	*UBE2D1*	ubiquitin-conjugating enzyme E2D 1
27	*POLR2A*	polymerase (RNA) II (DNA directed) polypeptide A, 220kDa
28	*FYN*	FYN proto-oncogene, Src family tyrosine kinase
29	*FBXO25*	F-box protein 25
30	*YBX1*	Y box binding protein 1
31	*FLNA*	filamin A, alpha
32	*UBQLN4*	ubiquilin 4
33	*THRAP3*	thyroid hormone receptor associated protein 3
34	*RAC1*	ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein Rac1)
35	*FHL2*	four and a half LIM domains 2
36	*EPS15*	epidermal growth factor receptor pathway substrate 15
37	*E2F1*	E2F transcription factor 1
38	*LRPPRC*	leucine-rich pentatricopeptide repeat containing
39	*PIK3R1*	phosphoinositide-3-kinase, regulatory subunit 1 (alpha)
40	*AHCYL1*	adenosylhomocysteinase-like 1
41	*PLK1*	polo-like kinase 1

Table 3.4. *List of hub targets with the gene name and description, ordered by centrality relevance.*

3.6 Breast cancer metastasis-free survival study

Gene signatures for predicting a breast cancer outcome are often used to stratify patients into prognostic groups. The members of the Wnt module as well as PPI hubs are considered to be candidate genes that confer an aggressive phenotype to breast cancer cells. Therefore, I was further interested in evaluating their prognostic power in the patient context of metastatic breast cancer. To that end, I first collected publicly available patient data.

3.6.1 Patient cohort and subtype prediction

Ten gene expression datasets of breast cancer patient samples were assembled into a compendium dataset. Annotations of metastasis events with available time to distant metastasis information were compiled together for 2075 patient samples. Within this cohort the molecular breast cancer subtypes were predicted using PAM50 signature, resulting in 1724 patients assigned with one of the following subtypes: basal-like (basal), HER2-enriched (Her2), luminal A (lumA) and luminal B (lumB) (Table 3.5, Figure 3.13A). As no sample was predicted as normal-breast-like subtype above the prediction strength threshold, this subtype was not considered in further analyses.

In the final dataset the 5-year metastasis-free survival (MFS) rate was highest for the lumA (0.92), whereas basal and Her2 subtypes had the lowest rate – 0.74 and 0.61, respectively. Furthermore, the predicted subtypes showed prognostic significance ($p = 2.37\text{E} - 12$) for the MFS (Figure 3.13B).

Subtype	N. of patients	N. of events	% of events	5-year MFS
Basal	289	74	26%	0.74
Her2	47	18	38%	0.61
LumA	716	116	16%	0.92
LumB	672	186	28%	0.76
Total:	1724	394	23%	0.82

Table 3.5. *Breast cancer patients with predicted subtypes: Table summarizes for the whole dataset as well as for each subtype separately the number (N.) of patients, the number of metastasis events, the percentage of the metastasis events within a group, and the 5-year metastasis-free survival (MFS) rate.*

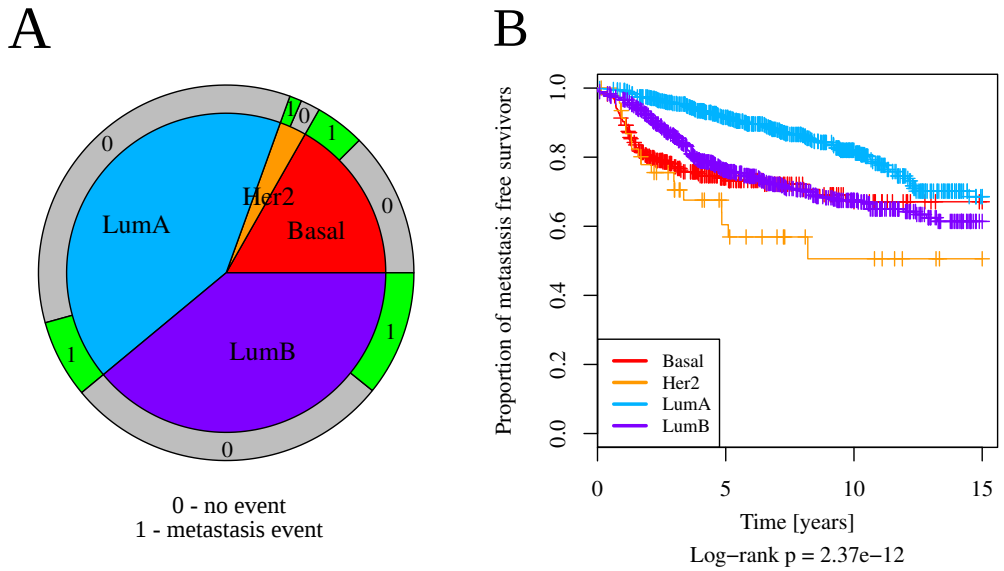

Figure 3.13. *Predicted breast cancer subtypes: (A) Pie chart depicts the proportion of subtypes for the 1724 patients. The outer ring represents the proportion of metastasis events in each subtype. (B) Kaplan-Meier curves showing MFS according to the four subtypes.*

3.6.2 Wnt module genes as prognostic signature

The members of the Wnt module were used as a pathway-based gene signature to test prognosis of metastasis development in breast cancer. First, expression levels of the 76 signature genes, which could be mapped to the expression data, were utilized for the correlation distance-based clustering analysis. Whereas the gene dendrogram revealed two major clusters, the patient dendrogram had a more complex structure (Figure 3.14).

The dynamic hybrid algorithm was used to automatically detect substructures within the patient dendrogram identifying four distinct clusters. These four clusters were significantly associated with different MFS (Figure 3.15). When exploring the composition of the subtypes across the dendrogram the cluster with the worst prognosis (pink cluster 3) showed a majority of basal but also smaller proportions of all three other subtypes. Interestingly, all clusters contained mixture of at least two or more subtypes.

Further, I inspected the positions of genes encoding the non-canonical Wnt ligands and receptors within the gene dendrogram. I found *WNT11*, *FZD5*

and *FZD4* within single gene cluster indicating similar co-expression patterns, whereas *WNT5A* and *ROR2* were separated from these within the second gene cluster (Figure 3.14).

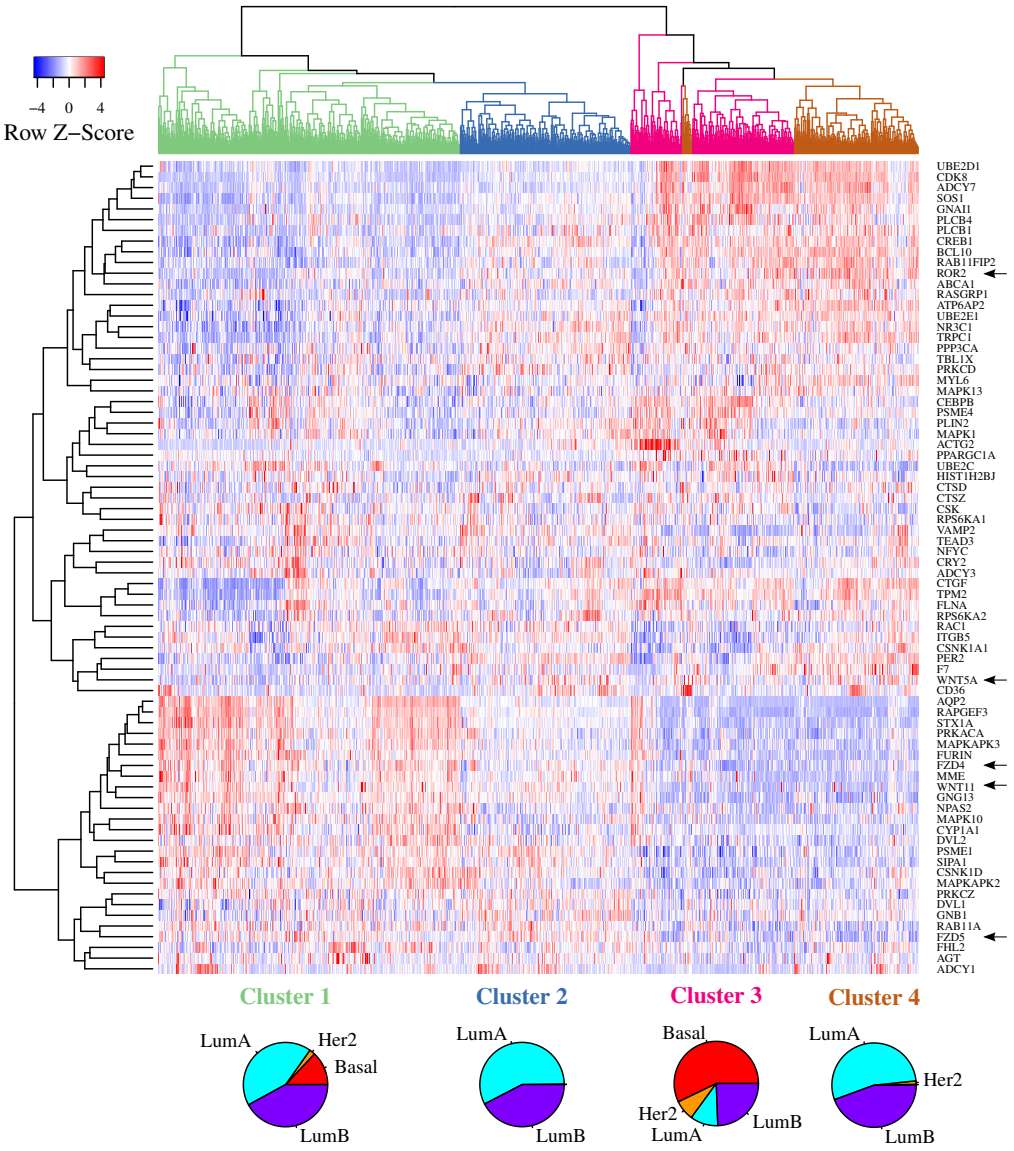

Figure 3.14. *The Wnt module gene signature in the whole patient cohort: Clustering analysis revealed the heatmap of expression levels (depicted as row z-scores) and the subsequent dendrogram shape-based analysis yielded four patient clusters. The pie charts at the bottom display subtype distribution within each cluster. The position of Wnt ligands and receptors on the gene dendrogram is depicted by the arrows.*

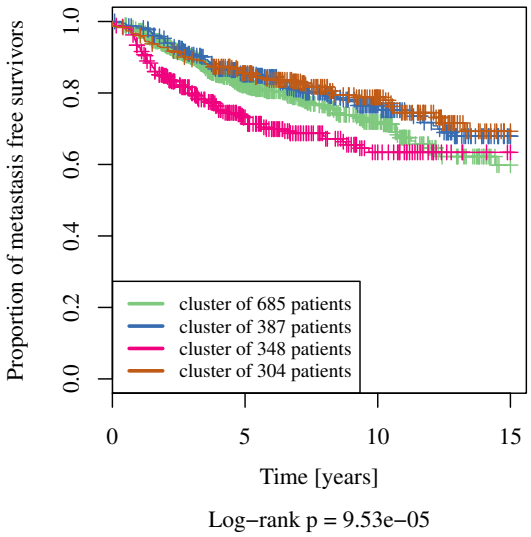

Figure 3.15. *The Wnt module gene signature in the whole patients cohort: Kaplan-Meier curves showing MFS according to the four clusters.*

Based on the mixed distribution of subtypes across the four patient clusters I hypothesized that there are different non-canonical Wnt pathway activation levels within the individual subtypes which can be associated with metastasis development. I focused on the lumA and basal subtypes and performed analogical signature-based clustering and Kaplan-Meier (KM) analyses as previously done for the whole cohort, but now in the subtype-specific context (Figure 3.16).

Clustering and cluster-detection analyses revealed two subgroups in the lumA subtype (Figure 3.16A), which further showed differences for MFS when tested in the KM survival analysis ($p = 0.0377$). The yellow cluster comprising 312 patients had better prognosis than the deep-sea-blue cluster of 404 samples. However, when considering Bonferroni multiple testing correction the difference was not significant ($q = 0.0754$). This can be explained by the proximity of the two KM curves for first 7-8 years, which only diverged from each other afterwards.

In the basal subtype-specific analysis two clusters were found based on the gene signature clustering (Figure 3.16B). They exhibited a significant difference for MFS ($p = 0.0145$, $q = 0.029$) with the KM curve of smaller (violet) cluster showing worse metastasis prognosis than the bigger (pea-green) cluster curve, especially within the first 5 years.

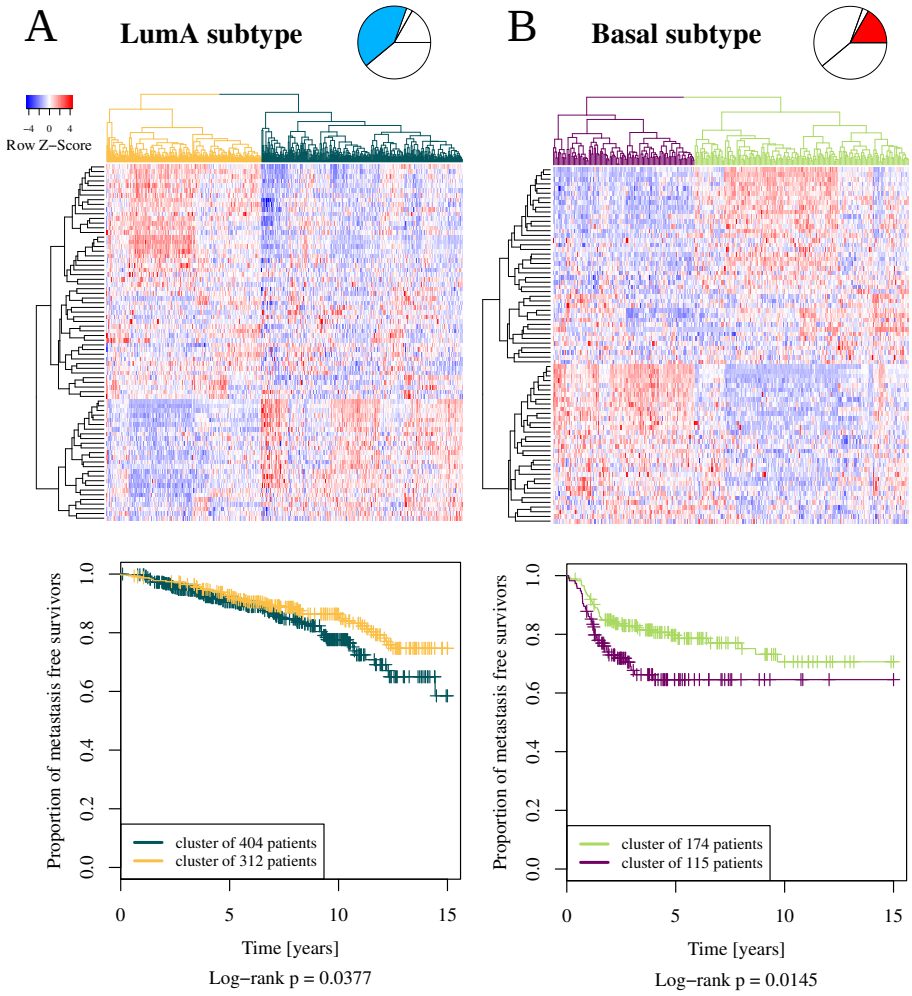

Figure 3.16. *Wnt module gene signature in (A) luminal A and (B) basal-like subtypes of breast cancer patients: The figure displays the heatmaps of expression levels with dendrograms from hierarchical clustering analyses at the top and the KM curves of the patient clusters from MFS analyses at the bottom.*

I was further interested in the discriminative potential of individual genes from the signature within the lumA and basal groups. Therefore, Cox regression models were fitted on the expression levels for each gene. The resulting values of hazard ration (HR) were used to define the *risk* genes (HR > 1) and the *protective* genes (HR < 1). The higher expression levels of risk and protective genes were associated with shorter and longer MFS, respectively. In the lumA group 12 genes were identified being associated with metastasis outcome at the p-value significance level $p < 0.05$: four risk genes and eight protective. Only one risk gene – *CSK*, c-Src tyrosine kinase – was detected at FDR < 0.05

level as significantly correlated with shorter MFS for lumA subtype patients. Within the basal patients nine and seven genes were found as potential risk and protective genes ($p < 0.05$), respectively. From there two passed FDR < 0.05 level: *FZD4*, frizzled receptor 4, as a risk gene and *PLCB1*, phospholipase C beta 1, as a protective gene. To visualized these results I mapped them back on the Wnt module topology (Figure 3.17) – although the topology itself was not exploited in the survival analysis.

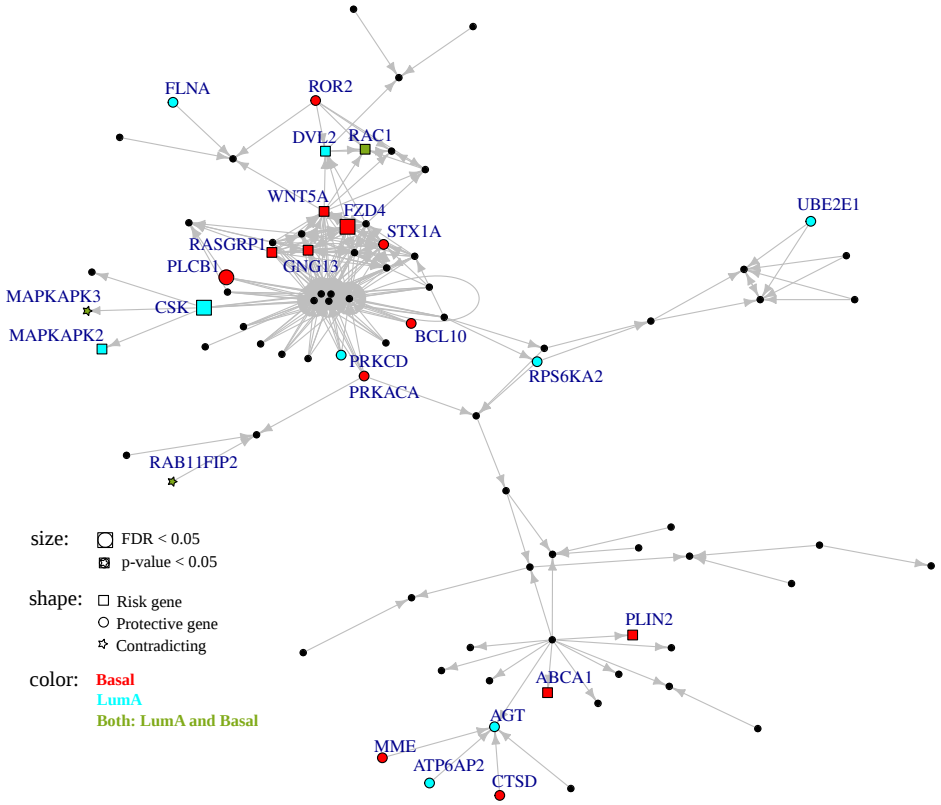

Figure 3.17. *The Wnt module with subtype-specific metastasis-free survival (MFS) associated genes. The size of nodes depicts the significance level. The color codes whether the gene was identified for basal patients (red), lumA patients (blue) or for both groups (green). The shape of nodes reflect a role of the genes: the risk genes associated with shorter survival (square), the protective genes with longer survival (circle) and genes with a contradicting role between the two subtypes (star). Small black nodes represent genes that were not associated with metastasis prognosis.*

3.6.3 PPI hub genes as prognostic signature

The gene signature comprising 41 hubs was applied to patient data in a similar fashion as the previous signature. Hierarchical clustering analysis was performed using 37 hub genes, which had representative gene probes in the expression data. Three clusters were detected in the patient dendrogram (Figure 3.18A). The clusters were subsequently subjected to MFS analysis, which revealed significant differences in prognosis between the cluster groups (Figure 3.18B). When investigating the subtype composition across the clusters I found all basal and Her2 patients grouped in the cluster 3 (pink) with the worst prognosis, whereas remaining two clusters contained mixture of lumA and lumB subtypes.

To assess a prognostic ability of individual hub genes within the whole patient cohort (not subtype-specific), I again utilized Cox regression models fitted on the expression levels of each hub. The analysis identified three protective and two risk genes that were significantly correlated with metastasis outcome on the FDR < 0.05 significance level (Appendix Table 5). The protective genes whose higher expression correlated with longer MFS were *EPS15*, *SFPQ* and *RPA2*. In contrast, the risk genes whose higher expression was associated with shorter MFS were *RAC1* and *PLK1*.

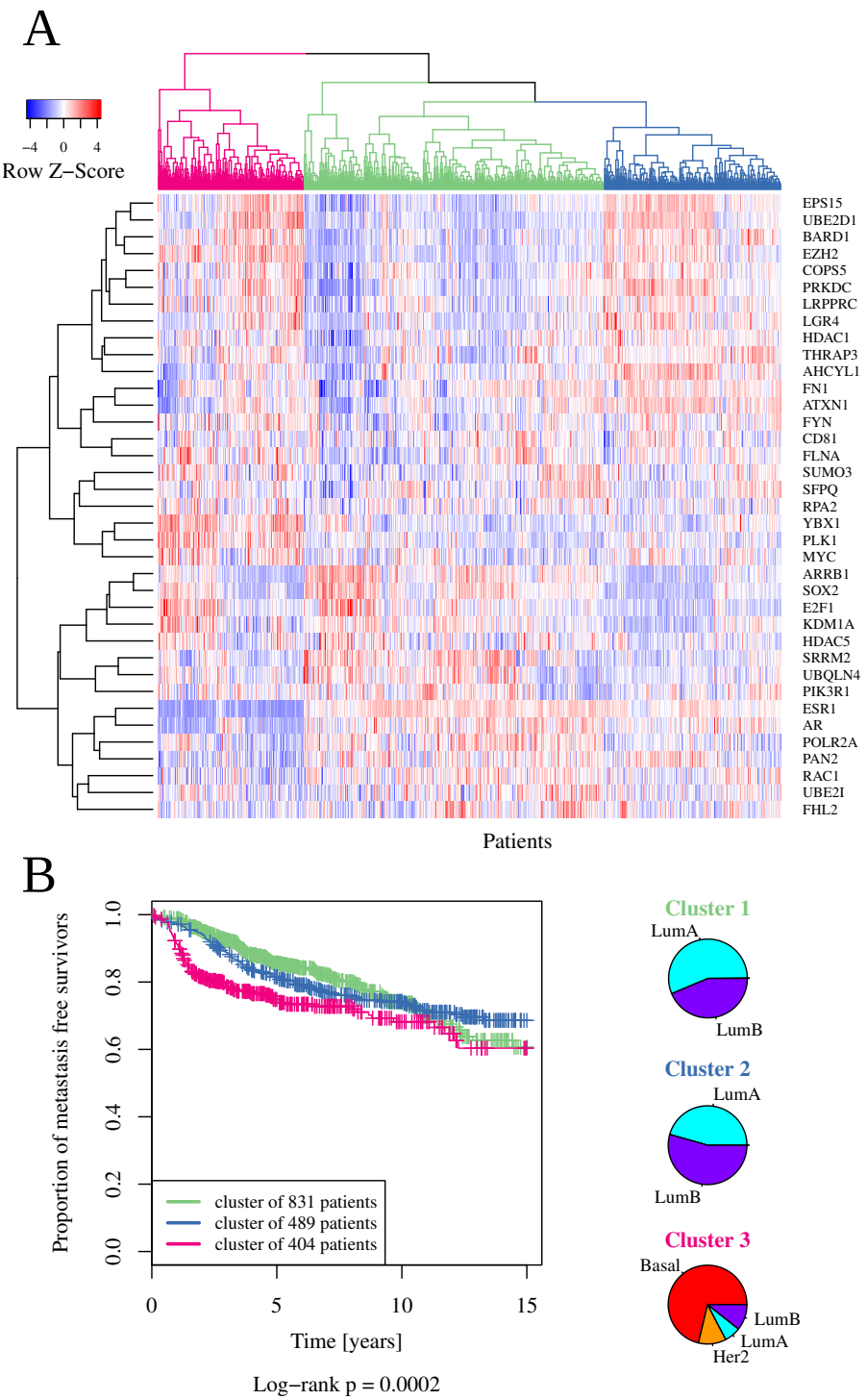

Figure 3.18. *The PPI hub genes as a prognostic signature in the whole cohort: (A) Clustering and subsequent dendrogram shape-based analyses yielded three patient clusters. (B) KM curves showed MFS according to the three detected clusters. The pie charts display the subtype distribution within each cluster.*

Discussion

4.1 Parsing and representing pathway knowledge

Existing pathway knowledge can greatly facilitate the interpretation of high-throughput data analysis results. Within this thesis I utilized pathway data for two main purposes: 1) to integrate them into enrichment methods for the comparative study and 2) to construct signaling networks of Wnt pathways.

A number of public databases offer signaling pathway data. In these databases pathway processes, like protein binding or signal transduction, are coded using different approaches (D'Eustachio, 2013). One of the current standards to encode pathway information is the BioPAX format. While the BioPAX definition itself is fixed (Demir et al., 2010), the specific syntax used and the level of detail vary between databases. I utilized the generic BioPAX parser *rBiopaxParser* developed by Kramer et al. (2013) to access pathway data and parse them into the R environment. In R the pathways are modeled as interaction graphs with the nodes representing molecules and the directed edges representing controlling processes such as activation and inhibition. Hence, knowledge on binding mechanisms or conversions of physical entities is not represented within these graphs.

I further transformed the graphs by removing non-protein-coding nodes in order to associate each node with a single gene symbol, and to preserve at the same time interaction continuity, which is important for signal propagation. Nevertheless, with only activation and inhibition processes being modeled, a lot of biological context is lost, which might be liable for multiple connected components being present in a single pathway graph (median of 2 components per KEGG pathway). Whereas this is not reflected heavily in the topology of Wnt networks as they were created by merging multiple pathways, it might

play a role in the subsequent performance of the PT-based enrichment methods (discussed further in the next section).

An alternative to *rBiopaxParser* is *graphite* R-package (Sales et al., 2012) offering pathways in a different representation – as graphs containing a mixture of directed and undirected edges. The directed edges represent positive or negative control mechanisms, whereas indirected edges stand for physical binding. Such a concept certainly models more details on pathway processes and the user can eventually select only the desired edges. However, the *rBiopaxParser* as a generic parser provides more flexibility in terms of generating and modifying pathway models.

Within the framework of the comparative study, I provided the same parsed pathway data for all methods to perform a fair comparison. Although all evaluated PT-based methods require some form of a directed graph as a pathway input, the specific representations of a pathway differ between the methods. For instance, the CePa methods operate on graphs whose nodes can comprise multiple molecules that represent a protein complex or a gene family. Therefore, the CePa functionality based on this feature was suppressed when I supplied the method with the customized pathway data, in which complexes were split into individual nodes.

In summary, the representation of pathway knowledge can range from simplistic gene sets, through directed graphs, up to sophisticated models with multiple types of nodes and edges. However, the choice of the optimal representation depends on the integration purpose, the experimental context and the analytical approach. Subsequently, this choice affects the final results of end-point analysis; however, it is often hard to quantify this effect.

4.2 Enrichment methods

Enrichment analysis is a widely used strategy to place the results of a differential analysis into the context of known groups of molecules related in the same biological functions and processes, such as signaling pathways. In this thesis I comparatively evaluated three GS methods against four enrichment methods that account for pathway topology structure.

So far, there has been very little work devoted to understand the contribution of pathway topology to the enrichment analysis results. Recently, Jaakkola and Elo (2015) evaluated a similar set of methods in an empirical study of 10 datasets, concluding that PT-based methods found more significant pathways than methods not using pathway topology. The authors highlighted two PT-based methods – SPIA and CePa – in terms of consistency of results across different data sets and the number of results. Although it might be challenging to assess ground truth information in real datasets, *more results* should not be straightforwardly considered to be *better results*.

In contrast to their work I evaluated the methods in two extensive simulation studies, accounting for different parameter configurations, as well as using a benchmark of 24 real datasets (Bayerlová et al., 2015a).

Within the simulations, several challenges arose. To capture a pathway deregulation in a graph representation and subsequently reflect this alteration in the synthetic expression data is a non-trivial task. I utilized three designs to allocate affected genes to the topological structure of a deregulated pathway: community, betweenness and neighborhood. Notably, in a certain setting a particular method might be favored due to the inherent algorithm properties, for instance, the CePa methods by the betweenness design and PathNet by the neighborhood design. Furthermore, biological signal is mediated through a pathway via complex mechanisms and it could be questioned how well these graph concepts reflect a real perturbation.

In the first simulation study, in which pathway input was represented by the original KEGG pathways, none of the PT-based methods showed outstanding performance – neither for any parameter configuration nor for the overall accuracy measure.

The original KEGG pathways exhibited considerable gene overlap among each other – a single gene was present in 2.4 pathways on average. When a pathway was called deregulated on a certain detection call (DC) level, it obtained corresponding number of genes assigned as affected. This resulted in a number of pathways that contained affected genes although they were not called deregulated. Consequently, these pathways were detected as false positives and led to very low specificity of all methods in Study 1. It can be assumed that these *accidentally* affected genes were allocated randomly in a pathway not

called deregulated, whereas in a pathway called deregulated they were placed in a topological design context. Therefore, the PT-based methods could be expected to deal with the problem of these false positives. For instance, the PathNet method claims to account for the pathway overlaps by constructing a pooled pathway. However, according to the results of this study the PT-based methods do not appear to solve this problem.

The problem of pathway overlaps has already been pointed out by several authors. To overcome this problem, Jiang and Gentleman (2007) suggested to test the subtracted gene sets or the intersects of gene sets. Later on, Tarca et al. (2012) introduced a method that down-weights overlapping genes in the enrichment analysis. However, these strategies were applied only to the GS methods.

Upon closer inspection of the results of Study 1, the CePa GSA method showed markedly different behavior than that exhibited by the other methods. On the one hand, its overall sensitivity was the highest, but on the other hand, its specificity was extremely low in the configurations, in which with a lot of genes were affected ($size = \{\text{big}\}$, $N = 70$ and $DC = 70\%$). Moreover, the CePa GSA was the best in identifying target pathways in the diseased benchmark datasets but it also identified more than half of all KEGG pathways as significant. This clearly distinct performance was expected from the single self-contained method included in this work. The self-contained methods already been proposed to be too powerful (Goeman and Bühlmann, 2007) and therefore lacking specificity. My results further indicate that the centrality-based gene weighting approach of the CePa GSA method does not improve this drawback.

The ranking-based functional class scoring (FCS) methods have already been reported to detect modest deregulation changes better then the ORA approach (Abatangelo et al., 2009). Results of the simulation Study 1 further suggest that the ranking tests (WRS and KS) are more sensitive than any of the PT-based methods when gene expression changes are weak (configuration with $mean = \pm 1$).

Simulation Study 2 was performed to compare the methods in a setting unbiased by false-positives originating from the pathway overlap. The real topology structures were preserved in the non-overlapping pathway input,

but unique synthetic IDs were generated for the graph nodes to prevent the overlap. The overall specificity increased in Study 2, especially, for the PT-based methods. In addition to the demonstrated high specificity, the four PT-based methods showed the highest overall accuracy.

Interestingly, KS reached the lowest overall specificity among all methods in Study 2, and specificity was always worse when more genes were affected (e.g. for configuration $size = \{\text{big}\}$, $N = 70$ and $DC = 70\%$). This might be explained by its null hypothesis, which actually tests whether a gene set or its complement is significantly enriched (Maciejewski, 2013). In the simulation configurations when the complement is enriched, false positive pathways are detected.

Although, the simulated data may not completely capture complex biology of gene expression, they provide controllable settings with defined results. In contrast to the simulations, the benchmark datasets represent real life settings. However, it is not possible to assess specificity and accuracy as the evaluated target pathway for each dataset is only one out of multiple potential true positive pathways. Nevertheless, in terms of identifying and prioritizing the target pathways the WRS and PathNet methods appeared to be reasonably successful. In conclusion, the overall results suggest that the GS approach might be satisfactory enough to detect significantly deregulated pathways.

In the enrichment analysis framework two additional aspects warrant further discussion: the pathway data – particularly their topological representation – and the experimental data that are used to test for pathway enrichment.

Pathway data stored in the databases are constantly curated and updated. It is easier to process and handle gene sets than to generate and integrate a topological pathway input of a new database version. Therefore, PT-based methods are less flexible in this respect and their utility is usually limited to their internal support of the processed pathway data. Furthermore, the true topology of a pathway is context-dependent and differs between organisms, cell types, and tissues (Khatri et al., 2012). Whereas the structure of a pathway in a database is usually given by a certain condition or state, the experimental data tested in the enrichment analysis comes from various biological systems.

Another issue concerning the enrichment analysis of pathways originates from gene expression levels being used as experimental data. Biological path-

ways are complex multi-layer systems consisting of membrane receptors, intracellular proteins, small molecules, transcription factors, and target genes. Although the transcriptome correlates with the proteome to 63% when considering relative abundances (Vogel and Marcotte, 2012), it has become common practice to use mRNA expression levels as proxies for abundances of corresponding proteins. Based on this approximation, differential expression of genes is utilized to detect altered pathways, even though signaling pathways are collections of mostly proteins. For gene sets this approximation assumes that if the mRNA levels are differential also corresponding proteins should have differential abundances. However, for the PT-based approach it implies further assumptions that might be incorrect, such as modeling activation and inhibition between proteins and assuming their specific place and role within a pathway structure based on the mRNA levels.

Therefore, this discrepancy between gene expression measurements and signaling pathway representations can result in a limited interpretability of the obtained results. Taken together, if a pathway gene set does not appropriately reflect the experimental data, it cannot be expected that adding topological information would improve the method's performance.

A simple solution to address this discrepancy would be to replace transcriptomic data by proteomic measurements for testing enrichment of signaling pathways. Technological advances in mass spectrometry and SILAC now allow to measure protein abundances on a large scale. Therefore, these data offer a reasonable alternative to mRNA levels as they reflect the protein character of the pathways much more closely.

Another straightforward solution was proposed by Naeem et al. (2012): To test sets of target genes which share a specific transcription factor. Within the analysis of gene expression data, testing such sets might be more appropriate than testing pathways. But *again*, this solution is applicable only within the GS approach.

In summary, whereas it is rather straightforward to supply a simple gene set of a pathway for GS methods, to represent the true topology of a pathway is more challenging. Moreover, testing the topological structure of a pathway using gene expression data implies assumptions which are not reflected in the biology of cellular signaling and gene expression responses. In general, more

comprehensive studies and further benchmark data are needed to systematically evaluate these methods in the context of various high-throughput technologies.

4.3 Wnt networks

In contrast to the arguable usage of pathway topology for testing an enrichment, describing a signaling pathway as a network structure using graph theory approaches represents a more promising concept. Wnt signaling pathways are of great interest due to their clinical importance as well as their ambiguous roles in breast cancer initiation and progression. I constructed four signaling networks, which were assembled from directed graphs of Wnt pathways from multiple databases, and reflect distinct Wnt signaling cascades: 1. canonical Wnt signaling, 2. non-canonical Wnt signaling, 3. inhibition of canonical Wnt signaling, and 4. regulation of Wnt signaling (Bayerlová et al., 2015b). These newly constructed Wnt networks represent a significant subset of the totality of machine-readable knowledge on Wnt signaling.

Network analysis approaches have proven to be powerful for elucidating important topological as well as functional elements of networks (Aittokallio and Schwikowski, 2006; Vidal et al., 2011). I utilized two approaches – centrality and network clustering – to explore the architecture of the Wnt networks.

First, I located central nodes reflecting essential topological elements. For each Wnt network I identified up to five top *key nodes*. These key nodes represent at the same time highly interconnected nodes (high degree hubs) as well as nodes with frequent signal flow (high betweenness) when considering directed signaling networks. The top key nodes in the canonical network were *CTNNB1*, *DVL1*, *DVL2*, *DVL3*, and *AXIN1*. While *CTNNB1* encoding β-catenin is the core member of the canonical Wnt signaling cascade, the others are also recognized for their essential role in mediating β-catenin-dependent Wnt signals (Moon et al., 2004). The key nodes of the non-canonical network comprised *WNT5A* and *CAMK2A* – well-known important components of β-catenin-independent signaling (Kühl et al., 2000) – and *TLN1* and *CALM3*. Identification of *TLN1* as a key node was particularly surprising, as this gene has not received a large amount of attention in the literature in the context of Wnt signaling. It encodes Talin-1, which is a major cytoskeletal protein

associated with integrin signaling and focal adhesion (Critchley and Gingras, 2008). the fact, that these processes are linked to cytoskeletal rearrangement, interestingly complements the role of the non-canonical Wnt pathway in cell migration. Within the "inhibition of canonical Wnt signaling" network, the key nodes included genes involved in the β-catenin destruction complex – *APC*, *AXIN1* and *CTNNB1* – as well as *GSK3B* and *CSNK1A1*, which are main negative modulators of canonical signaling (Katoh and Katoh, 2006). In the "regulation of Wnt signaling" network, the top key nodes were upstream regulators of Wnt signaling: Hedgehog pathway members *GLI1*, *GLI2*, and *GLI3* which suppress Wnt signals (He et al., 2006) and the *MYC* gene encoding the c-myc protein known for activating canonical Wnt signaling (Cowling et al., 2007).

The second utilized network approach – network clustering – focused on the selected subset of frequently reviewed Wnt pathway members (Moon et al., 2004; Katoh and Katoh, 2007; Amerongen and Nusse, 2009; Holland et al., 2013). The identified subnetwork of well-established genes in Wnt signaling was subjected to community detection analysis. Five topological modules were revealed as densely interconnected communities (see Figure 3.7) and could be denoted as the ligand layer (purple), the PCP (yellow) and Wnt/Ca^{2+} (green) branches, the signal transdution core (blue) and the effector layer (red). Biological pathways in general were reported to be not modular and rather to represent individual closely connected graphs (Kirouac et al., 2012). Nevertheless, the Wnt subnetwork could be partitioned into biologically meaningful modules representing either alternative Wnt cascades or different segments of signaling layers.

In summary, the identified key nodes fall in line with the hallmarks of Wnt signaling reported in literature. Besides underlining the importance of known essential Wnt signal mediators, the key nodes also revealed *TLN1*, a gene less known for its role in Wnt context creating thus opportunities for further experimental investigation. Furthermore, the topological modules identified in the community detection analysis reliably reflect different functional blocks of Wnt signal transduction.

Nevertheless, there are several limitations concerning the features and the usage of the networks, three rather general and two specific ones for the new Wnt networks:

1. Despite the growing amount of data in pathway databases, the currently available pathway data are still incomplete and can contain biases. Given the dependency of the Wnt networks on database knowledge, it is likely that they inherited these limitations.
2. Pathways in the databases were derived under varying conditions. Therefore, the constructed Wnt networks represent a set of possible interactions that might occur between the molecules ignoring a specific context.
3. Generally speaking, networks as models capturing cellular pathway structure can be conceptually useful. However, it might be challenging to quantify to what extent they actually resemble physical reality, which may be to a rather small extent (Ideker and Krogan, 2012).
4. The network construction step concerning the stratification of collected pathways into the four conceptual groups was based on literature and expert knowledge, and humans are known to lack objectivity.
5. As already pointed out previously in section 4.1 *Parsing and representing pathway knowledge*, the edges of activation and inhibition controlling processes represent only a part of signaling mechanisms. Further interaction types are not included in the Wnt networks.

Therefore, future improvements of the Wnt networks could be achieved by integration of PPI information to gain a more comprehensive picture of signaling networks. However, it is worth noting that increased complexity also demands more cautious analysis and interpretation.

From the point of functional reduction of network complexity, gene expression data can be integrated with networks in order to extract relevant information and to ease interpretation of biological phenomena underlying experimental data.

4.4 Targets and network integration

Pathway interventions at multiple levels can reveal particular signaling mechanisms and their consequences. Activation states of distinct Wnt signaling pathways in breast cancer cells, in particular the canonical and the non-canonical Wnt cascades, have so far eluded detection. Based on the work of Klemm et al. (2011) I hypothesized that the non-canonical WNT pathway is critical for cancer cell proliferation and migration. Following this hypothesis the MCF-7 cells were perturbed by the non-canonical receptor Ror2 and ligand Wnt5a: In particular, the cells were transfected with Ror2 over-expression construct and/or stimulated with recombinant Wnt5a. The phenotype of the estrogen receptor positive MCF-7 cell line is considered to be weakly invasive corresponding to lumA breast cancer subtype. The major consequence of the individual as well as combined perturbations at the phenotypic level was increased cell invasion.

The RNA of the cells was deep-sequenced to explore the consequences at gene expression regulation level by identifying differential targets. Although the Wnt5a stimulated cells showed enhanced invasiveness, the numbers of targets were moderate (seven for ctl vs. ctl + Wnt5a and eleven for Ror2 vs. Ror2 + Wnt5a comparisons). Certainly, the most interesting differentially expressed target of Wnt5a stimulation is the *ROR2* gene which was upregulated in the ctl + Wnt5a samples compared to control samples. This positive-feedback loop supports the theory that Ror2 acts as a receptor for Wnt5a ligand (Oishi et al., 2003). The only common differential gene in both comparisons testing for Wnt5a stimulation effect in MCF-7 was *MUC5AC*, mucin 5AC. It has been studied in the context of colorectal (Walsh et al., 2013) and pancreatic cancer (Hoshi et al., 2011), and in the latter cancer type its expression was associated with tumor growth. However, to my best knowledge, so far it was not linked with invasive breast cancer neither reported as a potential target of Wnt5a signaling.

In contrast to a few target genes affected by Wnt5a stimulation, the three comparisons testing for the Ror2 over-expression effect revealed 2068 common DEGs representing stable targets of this perturbation. These common targets of Ror2 over-expression are considered to be candidate genes that confer the invasive phenotype to MCF-7 breast cancer cells. To gain further insight

into the biology underlying this fairly long list of expression-responsive targets I took three bioinformatic approaches: **1.** enrichment analysis of gene sets originating from the Wnt networks (results of 3.4.1 section), **2.** non-canonical Wnt network integration with the targets (results of 3.5.1 section), and **3.** identification of hub targets in PPI network (results of 3.5.2 section).

1. For enrichment testing the over-representation analysis (ORA) utilizing the 2×2 table-based Fisher's exact test was chosen because of the given cut-off setting. PT-based methods were not considered because of the issues discussed in 4.2 *Enrichment methods* section. Despite the obvious mRNA versus protein approximation (also discussed in 4.2), the ORA detected the gene set of the non-canonical Wnt signaling pathway as significant in the target list as well as in the sublist of only up-regulated targets. It suggests that activation of non-canonical Wnt signaling was induced by the Ror2 over-expression and this activation state can be linked with the increased invasiveness. Besides the non-canonical Wnt gene set, the "regulation of Wnt signaling" gene set was significantly over-represented in the target list, which indicates that pathways acting upstream Wnt were also activated by Ror2 over-expression.

2. Further, the non-canonical Wnt network was used for target integration as it proved to be significant when utilized as a gene set. This network is a signaling network of 489 nodes representing intermediate size compared to large-scale interactomes. Therefore, I have not considered elaborated methods designed for an active-module-search in large networks – e.g. methods of Ideker et al. (2002); Dittrich et al. (2008), and Ma et al. (2011). Instead, I have chosen rather straightforward approach of a direct projection of the targets onto the Wnt network nodes combined with Steiner tree analysis. In this way I identified the differentially regulated module of the non-canonical Wnt pathway responsive to Ror2 over-expression. Into this module Steiner nodes were introduced. Although the genes of Steiner nodes themselves were not detected as differentially expressed, they represent important connectors of the targets.

By means of differential expression analysis, key cancer drivers, such as Her2, p53, and KRAS, are typically not detected; however, they play a central role in networks by interconnecting those genes, which are expression-responsive (Chuang et al., 2007). Therefore, I have focused on the identified

Steiner node genes. Several of these genes have already been investigated in the context of aggressive breast cancer, for instance: *CD36*, *PPARGC1A*, and *CSNK1D*, as well as well-known *WNT5A* and *DVL2*. Signaling activated by Wnt5a and mediated via Dvl2 was previously shown to promote migration of breast cancer cells (Zhu et al., 2012), although the evidence on the role of Wnt5a signaling in breast cancer progression remains ambiguous (McDonald and Silver, 2009). The expression of *CD36* was reported to be very low in highly aggressive MDA-MB-231 breast cancer cells compared to relatively non-proliferative cell lines (Uray et al., 2004). This is corresponding to its (although non-significant) down-regulation in the Ror2 over-expression conditions with the increased cell invasiveness. The *PPARGC1A* gene encodes a transcriptional regulator of multiple metabolic pathways and its increased expression was associated with the ability of breast cancer cells to metastasize (Chen et al., 2007; Bhalla et al., 2011). Similarly, the expression of *CSNK1D* was associated with lymph-node-positive breast carcinomas (Abba et al., 2007).

In summary, the Steiner nodes in the context-specific non-canonical Wnt module are highlighted as topologically-essential elements as well as genes with functional relevance for breast cancer progression. Nevertheless, it should be noted that this integration approach by definition is unable to identify any new interactions leading to network re-wiring.

3. By identification of hub targets in the PPI network, I aimed to explore functionally interconnected targets that operate outside the Wnt signaling pathway. As breast cancer progression is a complex mechanism linked with changes in multiple processes, the hub targets can point towards key players within these processes. Network hubs in general were reported not only as central interactive nodes but also as having an essential biological role within the networks (Barabási et al., 2011).

The identified hubs have to be considered as context specific, meaning that there might be more central nodes within the whole PPI network; however, the identified hubs reflect central nodes only within PPIs of the targets. When interpreting the hubs the research and database biases should be taken into account. In the case of proteins, which have been the subject of particular focus in research, there are more relevant results and database entries. Therefore, the target hubs should not only be seen as biologically important but also could be

simply *intensively-studied*. This is illustrated on the examples of *ESR1*, *AR* and *MYC* hubs, which are extensively discussed in literature in breast cancer context (Yager and Davidson, 2006; Fioretti et al., 2014; Watson et al., 1991).

The target hubs overlapping with the differentially regulated Wnt module genes return the attention back to Wnt signaling. The overlap points out less renowned Wnt network members such as *UBE2D1*, *FLNA*, and *FHL2* but also the well-recognized *RAC1* gene.

In summary, the results of these three analysis approaches highlight the importance of non-canonical Wnt signaling, as well as reveal further key genes in aggressive breast cancer. Nevertheless, to establish a more comprehensive picture of the role of target hubs and non-canonical Wnt module genes in breast cancer, I further translated these results into clinical context of metastatic breast cancer patients.

4.5 Breast cancer metastasis and prognostic genes

The Ror2 over-expression targets are considered to be genes implicated in the increased invasion of MCF-7 breast cancer cells. From these 2068 targets I selected two moderately short, context-specific subsets: 84 non-canonical Wnt module genes and 41 hub genes, using network integration approaches. To validate their role in breast cancer progression, I utilized these two subsets as prognostic signatures for the metastasis-free survival (MFS) analysis of breast cancer patient cohort.

To this end, the expression profiles of patients were collected across ten public datasets. Within this cohort the breast cancer molecular subtypes were predicted using *pam50* signature. Recently, Patil et al. (2015) showed that a prediction of particular sample class based on a gene signature can change depending on the background represented by the rest of samples in the dataset. Therefore, the predicted classification based on the signature should be treated with caution.

To check the quality of the predicted subtypes stratification at a basic biological level, I investigated them in the MFS analysis. The results are consistent with the relapse-free survival observed in the study of Parker et al. (2009), which confirms the association of breast cancer subtypes with different

metastatic potentials. Similarly, according to 5-year MFS the lumA subtype showed the best prognosis, as expected (Millar et al., 2009), whereas the Her2 subtype was found to show the worst prognosis, followed by the basal subtype.

The genes of the non-canonical Wnt module used as a signature for clustering analysis of this cohort revealed patient subgroups with different prognosis. The four subgroups appear to have two main expression pattern suggesting two modes of non-canonical Wnt pathway activation. The two modes are represented in the gene dendrogram by two clusters: Whereas one cluster includes *WNT5A* and *ROR2*, the second contains *WNT11*, *FZD5* and *FZD4* as closely correlated genes. These similar co-expression patterns of the genes can be interpreted as an indication of similar functions or involvement in the same processes (Allocco et al., 2004). The ligand-receptor role of Wnt5a-Ror2 in breast cancer has been previously reported (Oishi et al., 2003) and was also indicated by the changes in Ror2 expression levels after Wnt5a stimulation in the RNA-Seq data. Moreover, Wnt11 and Fzd4 were suggested as a ligand-receptor couple in the context of cardiomyocyte differentiation (Abdul-Ghani et al., 2011) and kidney development (Ye et al., 2011); however, this link has not been discussed so far in the context of breast cancer.

The patient cluster with the poorest prognosis was found to contain a large proportion of the basal subtype patients, which is consistent with an increased likelihood of metastasis development in the triple-negative breast cancer (Dent et al., 2007). However, each subtype was distributed across two or more clusters. This indicates that the clusters do not fully reflect the biology underlying breast cancer subtypes and that there are differences in the expression of non-canonical Wnt module genes also within the particular subtypes.

Therefore, I speculate that the differences in metastasis development may be associated with varying levels of non-canonical Wnt pathway activation within the individual subtypes. I focused on the lumA and basal groups as representatives of generally good and bad prognosis, respectively. Her2 subtype was not considered due to the small sample size.

Although lumA tumors have been repeatedly observed to have the most favorable prognosis in terms of MFS (Sørlie et al., 2003; Millar et al., 2009; Voduc et al., 2010), there are lumA cases that relapse and metastasize. It is of high clinical interest to identify low-risk lumA patients who might be spared

breast radiotherapy. The signature of non-canonical Wnt module genes applied to lumA patients revealed two subgroups; however, the difference between the subgroups in the MFS was not significant after the Bonferroni correction. In contrast to the aggressive subtypes, which tend to relapse early within the first 5 years, the luminal groups are characterized by continued relapses between 5 and 15 years (Kennecke et al., 2010). Therefore, the lack of statistical significance could originate from the proximity of the two KM curves within the initial years.

Within the basal subtype, further subgroups were previously identified and linked to prominent biological and transcriptional differences (Lehmann et al., 2011; Neve et al., 2006). I demonstrated that two clear basal subgroups with distinct prognosis of metastasis development can be defined based on the expression of non-canonical Wnt module genes. In summary, these results indicate a substantial role of non-canonical Wnt signaling mediated via the Ror2 receptor in the development of metastasis from both basal and lumA primary tumors.

Further, I attempted to identify single gene markers from the Wnt module that correlate with metastasis outcome in the two subtypes. In the lumA subtype, higher expression levels of Steiner node gene *CSK*, c-Src tyrosine kinase, were associated with shorter metastasis-free survival. This result is in line with previous reports that the activation of c-Src is critical for the progression and metastases of ER-positive breast cancer (Planas-Silva et al., 2006). For the basal patients, increased expression of *FZD4* was revealed as a strong risk factor for metastasis development. Conversely, higher levels of *PLCB1* were found to predict longer metastasis-free survival for this breast cancer subtype. However, neither *FZD4* nor *PLCB1* could be found previously discussed in the literature in the context of subtype-specific breast cancer progression.

Several additional genes were detected on a lower significance level ($p < 0.05$) as prognostic for each subtype. Of these genes, it is noteworthy that *WNT5A* was identified as a risk gene for the basal subtype, which contributes a new piece of evidence to its otherwise poorly understood role in breast cancer. Interestingly, when the prognostic genes are mapped back on the module topology, most of the the basal-associated prognostic genes tend to cluster

within the central core of the module, whereas lumA-associated genes appear to be distributed across the whole module structure. This pattern, combined with the little overlap exhibited between lumA and basal prognostic genes, highlight the diversity of these two subtypes. Moreover, it points towards different metastatic mechanisms being used by each subtype even within the context of a single signaling pathway.

To assess the clinical importance of the hub targets similar analysis approaches were applied. Based on the expression profiles of the target hub genes, three breast cancer patient groups can be defined. The group with the poorest prognosis clusters all basal and Her2 patients, which is not unexpected considering the presence of *ESR1*, estrogen receptor 1, in the hub signature.

The hub targets that were identified as *protective* genes, whose higher expression was linked to favorable prognosis, comprise *EPS15*, *SFPQ* and *RPA2*. The detection of *EPS15* as a protective gene is in line with the results of a study by Dai et al. (2015). Dai and colleagues conducted analysis of six expression datasets and reported that the over-expression of *EPS15* was a favorable prognostic factor for overall survival, especially in tumors with positive ER status. In this study I extend this finding to prognosis in terms of metastases, although I have not further explored the patient cohort for the *EPS15* prognostic power within the particular subtypes.

Surprisingly, the estrogen receptor *ESR1* had only moderately decreased hazard ratio (HR = 0.929), and its expression was not identified as significantly associated with favorable prognosis. This could be caused by the selection of a single microarray probe to represent a gene expression and it raises the question of whether the probe with the highest average expression level for the gene is the most *representative*. Miller et al. (2011) identified the probe with the highest average expression as best choice for between-study consistency. Nevertheless, further studies would be desirable to evaluate optimal strategies for aggregating multiple gene probes within compendium datasets.

The risk prognostic factors detected within the hub targets includes *RAC1* and *PLK1*. Both genes are considered to be promising therapeutic targets whose inhibition can block breast tumor growth (Cardama et al., 2014; Castillo-Pichardo et al., 2014; Hu et al., 2012; Yao et al., 2012; Bhola et al., 2015).

Although network hubs are generally regarded as biologically essential, Yang et al. (2014) reported that prognostic genes in different cancer types – including breast cancer – tend not to be hub genes. However, this conclusion was drawn based on co-expression network evaluation.

A more general criticism of using gene expression signatures was raised by Venet et al. (2011), who showed that random gene signatures are also associated with breast cancer outcome. They explain this by the fact that several members of such random signatures are likely to be associated with proliferation, and that proliferation genes define breast cancer transcriptome into large extent.

However, within this study, the application of gene signatures is not primarily intended for establishing new clinical subgroups of patients. Rather, the aim pf the study is to validate the implication of the Wnt module and the target hubs – two context-specific subsets of Ror2 over-expression targets – in cancer progression. Nevertheless, several individual genes from both signatures could be significantly associated with breast cancer outcome, suggesting them as powerful indicators of breast cancer prognosis.

Conclusions and Outlook

Analyzing complex data such as gene expression profiles is challenging, therefore integration of prior biological knowledge is crucial for a successful interpretation of the results. Signaling pathways and molecular networks represent such a source of external information, which allows for the embedding of results of differential analysis within a biological context. Various methods within pathway enrichment and network analysis approaches have been designed for this integration purpose.

The integration of existing knowledge underlies both major questions (see 1.7 *Aims and organization of the thesis*), which defined the scope of this thesis to a large extent. In order to answer the first question:

- What are benefits and costs of integrating pathway-topology information into enrichment analysis?

I comparatively evaluated three gene-set (GS) methods against four pathway topology-based (PT-based) methods. These methods were compared on the same pathway data in two extensive simulation studies accounting for different parameter configurations and on a benchmark of 24 real datasets. In the benchmark data analysis, both types of methods showed a comparable ability to detect enriched pathways. The PT-based methods showed better performance in the simulation scenarios with non-overlapping pathways; however, they were not conclusively better in the simulation scenarios with realistic pathways exhibiting overlaps. When considering the effort involved in processing and integrating the customized pathway data, the GS methods are favorable. In summary, the simple GS enrichment approach appears satisfactory to detect significantly altered pathways.

Furthermore, I pointed out the discrepancy between gene expression measurements and the topological representation of a signaling pathway used for

enrichment testing. For a meaningful interpretation of pathway enrichment using its structure, it is necessary to accommodate experimental data and prior knowledge into the same conceptual cellular layer. Moreover, dealing with overlapping pathways appears challenging for both types of methods. Therefore, methodological improvements are needed in order to deal with the problem of pathway overlap. Also, further effort should be dedicated to understand the contribution of pathway topology in enrichment analysis.

Nevertheless, this is one of the first comprehensive comparative studies conducted on evaluation of PT-based enrichment methods. These results, as well as the discussion of strengths and limitations of the methods, may stimulate further scientific research in this field.

The second principal question within this thesis was stated as:

- Which module of the Wnt signaling network is active in aggressive breast cancer?

To answer this question, first, signaling networks that represent the distinct Wnt pathways were constructed. To elucidate important topological elements of these networks, I identified their key node genes, which were consistent with the hallmarks of Wnt signaling reported in literature. Furthermore, the subnetwork of well-established selected members of Wnt signaling was extracted from these networks. Its structure revealed five densely interconnected communities, which reliably reflect different functional blocks of Wnt signal transduction. This indicates that topological modules also have the potential to harbor functionally related genes. Nevertheless, the newly constructed Wnt networks have also some limitations. These limitations are inherited from the pathway database data as well as given by the level of modeled details and could be improved in the future by integration with further types of interactions.

These Wnt networks allowed further investigation of the role of Wnt signaling pathway in breast cancer cells. The initial hypothesis stated that non-canonical Wnt signaling pathway is critical for cancer cell proliferation and migration. Over-expression of the non-canonical Wnt receptor Ror2 in weakly invasive breast cancer cells resulted in increased cell invasion. Differentially expressed targets of this perturbation were enriched in the gene set of the non-canonical Wnt network; however, not in the canonical Wnt gene

set, underlining the importance of non-canonical Wnt signaling in aggressive breast cancer. Subsequently, in the network integration analysis, I identified a differentially regulated module of the non-canonical Wnt pathway, and highlighted its topologically-essential elements as well as genes functionally relevant for breast cancer progression, such as *WNT5A*, *DVL2*, *CD36*, *PPARGC1A*, and *CSNK1D*. To further scrutinize the genes within differential targets outside of Wnt signaling I utilized a protein-protein interaction (PPI) network. The identified PPI hubs represent highly interactive nodes, which highlight operationally important and/or intensively-studied targets.

Thus, I demonstrated that the complexity of molecular networks and of expression data can be reduced by mutual integration in order to extract functional information and to understand underlying biological phenomena.

As the Ror2 over-expression targets have been implicated in the invasiveness of breast cancer cells, I further validated the role of the non-canonical Wnt module genes and the target hubs in the gene expression data of metastatic breast cancer patients. These two subsets of targets were applied as gene signatures to cluster the patient cohort and demonstrated prognostic potential in terms of metastasis-free survival (MFS). The co-expression patterns of Wnt5a-Ror2 and Wnt11-Fzd4 across patient profiles indicate their ligand-receptor relation in breast cancer which was supported by literature in a similar context.

Furthermore, MFS analysis results indicated a substantial role of the non-canonical Wnt module genes in the development of metastasis within luminal A (lumA) and basal-like (basal) molecular breast cancer subtypes. Moreover, from the individual genes *CSK* was linked with favorable prognosis in the lumA patients, whereas *FZD4* and *PLCB1* were found to be risk and protective genes, respectively, in the basal subtype. Also, multiple hub targets could be associated with breast cancer outcome, including genes which were already reported as potential therapeutic targets, such as *RAC1* and *PLK1*.

Although signaling and its consequences mediated by particular Wnt ligand-receptor coupling remain incompletely characterized, the deregulation of the non-canonical Wnt pathway by Ror2 over-expression was shown to contribute to the clinical outcome in breast cancer. Therefore, these results pave the

way towards targeted therapies aimed at interfering with non-canonical Wnt signaling.

In conclusion, bioinformatic approaches derived from graph theory concepts as well as enrichment analysis were demonstrated to be powerful tools for deciphering complex gene expression patterns. Appropriate integration of breast cancer expression profiles with pathway and network data provides valuable insights into biological processes underlying breast cancer phenotypes.

Appendix

Table 1. *Summary of 26 parsed pathways from five public database used for the Wnt networks construction. The pathways were sorted based on literature and expert knowledge and each pathway was assigned into one of the group for network construction. In case a single interaction graph of a pathway contained multiple connected components (ConnComp), the components were also inspected for the network group membership and could be assigned into distinct groups.*

Assigned network	Database	Pathway name	Number of nodes
Canonical WNT signaling	BioCarta	wnt signaling pathway	27
Canonical WNT signaling	BioCarta	inactivation of gsk3 by akt causes accumulation of b-catenin in alveolar macrophages	25
Canonical WNT signaling	BioCarta	segmentation clock	24
Canonical WNT signaling	BioCarta	multi-step regulation of transcription by pitx2	25
Canonical WNT signaling	PID	Wnt signaling network	5
Canonical WNT signaling	PID	Canonical Wnt signaling pathway	18
Canonical WNT signaling	PID	Regulation of nuclear beta catenin signaling and target gene transcription	60
Canonical WNT signaling	Reactome	WNT ligand biogenesis and trafficking	21
Canonical WNT signaling	Reactome	TCF dependent signaling in response to WNT	155
Canonical WNT signaling	Reactome	WNT mediated activation of DVL	6
Canonical WNT signaling	Reactome	formation of the beta-catenin:TCF transactivating complex	45
Canonical WNT signaling	KEGG	Wnt signaling pathway	ConnComp of 117
Non-canonical WNT signaling	PID	Noncanonical Wnt signaling pathway	30
Non-canonical WNT signaling	Reactome	WNT ligand biogenesis and trafficking	21
Non-canonical WNT signaling	Reactome	WNT5A-dependent internalization of FZD2, FZD5 and ROR2	13
Non-canonical WNT signaling	Reactome	PCP/CE pathway	79
Non-canonical WNT signaling	Reactome	beta-catenin independent WNT signaling	501
Non-canonical WNT signaling	KEGG	Wnt signaling pathway	ConnComp of 23
Inhibition of Canonical WNT signaling	Pathway Commons	Beta-catenin phosphorylation cascade	62
Inhibition of Canonical WNT signaling	Pathway Commons	Degradation of beta-catenin by the destruction complex	51
Inhibition of Canonical WNT signaling	PID	Degradation of beta catenin	17
Inhibition of Canonical WNT signaling	Reactome	deactivation of the beta-catenin transactivating complex	9
Regulation of WNT signaling	PID	Presenilin action in Notch and Wnt signaling	40
Regulation of WNT signaling	PID	Validated targets of C-MYC transcriptional repression	71
Regulation of WNT signaling	KEGG	Basal cell carcinoma	46
Regulation of WNT signaling	KEGG	Hedgehog signaling pathway	55

Table 2. *Differential targets of Wnt5a stimulation in the two comparisons.*

Control versus Wnt5a stimulation			
Symbol	**Gene type**	**logFC**	**FDR**
ROR2	protein coding	3.51	5.1E-014
RP11-504P24.6	3prime overlapping ncrna	-9.67	5.9E-012
RP11-504P24.4	lincRNA	2.78	8.31E-07
RP11-3P17.5	lincRNA	1.42	0.014
SEMA3B	processed transcript	0.44	0.014
MUC5AC	protein coding	0.55	0.043
RP13-996F3.3	pseudogene	0.89	0.043
Ror2 versus Ror2 + Wnt5a stimulation			
Symbol	**Gene type**	**logFC**	**FDR**
AQP3	protein coding	0.54	0.0008
S100A14	protein coding	0.39	0.0008
PGR	protein coding	-0.56	0.009
MUC5AC	protein coding	0.75	0.009
AGR3	protein coding	-0.51	0.015
TIMP3	protein coding	0.46	0.019
DUSP4	protein coding	0.41	0.019
KRT7	protein coding	0.55	0.022
VIL1	protein coding	0.47	0.027
SNORD3A	lincRNA	-1.8	0.030
GCOM1	protein coding	0.85	0.033

Table 3. *List of the nodes of non-canonical Wnt module: measured in RNA-Seq data of cell lines with average fold-changes (logFC) from the three differential comparisons and in patient luminal A and basal-like subtypes with hazard ration (HR), p-value (p-val), and false-discovery rate (FDR). "NA" is introduced when the gene was not measured in the RNA-Seq or in patient microarrays.*

Node	Cell lines	LumA patients			Basal patients		
	logFC	HR	p-val	FDR	HR	p-val	FDR
PRKCZ	-0.25	0.404	0.125	0.419	1.642	0.503	0.759
FLNA	0.502	0.341	0.011	0.174	0.89	0.761	0.912
PSME1	0.204	0.466	0.098	0.419	0.661	0.455	0.743
PSME4	0.269	0.517	0.21	0.572	0.976	0.964	0.964
CSK	-0.065	10.111	0.001	0.04	0.391	0.233	0.524
RAB11A	-0.195	0.943	0.899	0.971	1.693	0.269	0.583
RAB11FIP2	0.366	0.26	0.044	0.313	6.077	0.045	0.226
FURIN	0.775	1.016	0.986	0.99	3.883	0.15	0.423
ATP6AP2	0.279	0.395	0.011	0.174	2.704	0.092	0.318
CTSZ	1.482	1.226	0.41	0.662	0.886	0.783	0.912
MME	1.001	2.346	0.06	0.35	0.234	0.007	0.085
TRPC1	0.571	1.608	0.501	0.705	3.18	0.278	0.587
TBL1X	0.282	1.004	0.99	0.99	0.595	0.222	0.524
UBE2E1	0.248	0.416	0.035	0.313	1.251	0.67	0.874
UBE2C	-0.252	1.089	0.831	0.943	2.102	0.089	0.318
UBE2D1	0.411	1.103	0.881	0.971	1.578	0.53	0.759
CDK8	0.306	2.215	0.123	0.419	2.682	0.09	0.318
MED30	-0.382	NA	NA	NA	NA	NA	NA
NR3C1	0.573	1.377	0.38	0.643	1.029	0.958	0.964
PRKACA	0.047	1.208	0.793	0.914	0.135	0.046	0.226
PLCB1	-0.281	0.752	0.374	0.643	0.08	0	0.001
RASGRP1	1.449	0.797	0.349	0.643	2.886	0.004	0.081
RAPGEF3	-0.526	0.201	0.091	0.419	1.511	0.716	0.892
CSNK1D	-0.169	0.782	0.616	0.768	1.877	0.347	0.642
SIPA1	0.408	0.817	0.794	0.914	1.103	0.931	0.956
ADCY3	0.247	0.296	0.136	0.419	0.889	0.881	0.92
ADCY7	0.919	0.681	0.528	0.716	0.633	0.525	0.759

Continued on next page

Table 3 – *Continued from previous page*

Node	Cell lines	LumA patients			Basal patients		
	logFC	**HR**	**p-val**	**FDR**	**HR**	**p-val**	**FDR**
ADCY1	-0.659	1.013	0.936	0.973	1.165	0.856	0.92
ADCY5	0.87	NA	NA	NA	NA	NA	NA
PRKCD	0.384	0.484	0.047	0.313	1.458	0.607	0.824
GNAI1	0.909	3.091	0.123	0.419	1.143	0.833	0.92
TEAD3	-0.275	1.323	0.673	0.825	2.367	0.44	0.743
CTSD	-0.293	1.359	0.138	0.419	0.307	0.026	0.163
PLCB4	-0.615	1.53	0.28	0.643	0.784	0.568	0.799
WNT5A	NA	1.19	0.46	0.686	3.649	0.004	0.081
WNT11	1.055	0.667	0.475	0.694	0.404	0.234	0.524
RAC1	-0.194	2.353	0.049	0.313	4.175	0.018	0.157
DVL2	-0.097	5.498	0.038	0.313	0.646	0.702	0.889
CSNK1A1	-0.285	1.863	0.137	0.419	1.293	0.637	0.849
MAPK10	0.044	0.736	0.584	0.765	0.819	0.792	0.912
DVL1	-0.273	1.072	0.907	0.971	5.908	0.052	0.235
ROR2	11.765	1.623	0.49	0.703	0.036	0.021	0.157
FZD5	0.323	4.242	0.208	0.572	0.148	0.306	0.609
PRICKLE1	-0.185	NA	NA	NA	NA	NA	NA
VAMP2	-0.382	0.781	0.719	0.868	0.753	0.804	0.912
ITGB5	0.561	0.701	0.341	0.643	0.912	0.866	0.92
PPP3CA	0.277	1.306	0.259	0.643	1.99	0.081	0.318
MYL6	0.275	0.68	0.46	0.686	2.399	0.15	0.423
TPM2	0.792	1.031	0.932	0.973	1.263	0.468	0.743
ACTG2	2.015	0.858	0.555	0.74	1.163	0.321	0.609
STX1A	0.538	3.406	0.073	0.397	0.036	0.014	0.151
GNB1	-0.279	2.627	0.098	0.419	1.197	0.779	0.912
GNG13	-0.727	2.307	0.37	0.643	44.671	0.048	0.226
FZD4	-0.668	0.59	0.389	0.643	34.826	0.001	0.02
MAPK13	0.473	0.889	0.79	0.914	1.218	0.679	0.874
MAPKAPK2	0.3	2.666	0.015	0.196	2.708	0.172	0.452
MAPKAPK3	0.291	0.139	0.003	0.105	7.857	0.019	0.157

Continued on next page

Table 3 – *Continued from previous page*

Node	Cell lines	LumA patients			Basal patients		
	logFC	**HR**	**p-val**	**FDR**	**HR**	**p-val**	**FDR**
CREB1	0.068	1.824	0.285	0.643	0.49	0.361	0.653
BCAR1	0.325	NA	NA	NA	NA	NA	NA
PER2	-0.005	0.645	0.273	0.643	2.156	0.21	0.515
AQP2	NA	0.389	0.284	0.643	3.322	0.393	0.678
F7	NA	1.526	0.211	0.572	13.421	0.11	0.363
AGT	NA	0.444	0.01	0.174	0.841	0.598	0.824
BCL10	-0.295	0.593	0.328	0.643	0.112	0.006	0.085
PPARGC1B	0.364	NA	NA	NA	NA	NA	NA
NFYC	-0.349	1.044	0.948	0.973	2.391	0.083	0.318
ABCA1	0.517	1.095	0.845	0.944	4.82	0.043	0.226
CD36	-0.207	0.854	0.369	0.643	0.629	0.145	0.423
CTGF	0.442	1.205	0.374	0.643	1.306	0.312	0.609
CYP1A1	2.101	1.758	0.606	0.767	5.751	0.207	0.515
FADS1	0.396	NA	NA	NA	NA	NA	NA
FHL2	0.998	0.881	0.516	0.713	0.898	0.758	0.912
GRHL1	-0.234	NA	NA	NA	NA	NA	NA
NPAS2	-0.033	0.641	0.389	0.643	0.506	0.309	0.609
PLIN2	0.543	0.776	0.445	0.686	2.22	0.023	0.158
PPARGC1A	NA	0.175	0.125	0.419	1.809	0.517	0.759
HIST1H2BJ	-0.475	0.297	0.352	0.643	3.958	0.369	0.653
H3F3B	-0.263	NA	NA	NA	NA	NA	NA
SOS1	0.31	0.259	0.114	0.419	1.586	0.469	0.743
RPS6KA1	-0.286	1.417	0.604	0.767	3.08	0.164	0.446
RPS6KA2	0.525	0.288	0.026	0.288	0.882	0.883	0.92
MAPK1	0.071	0.66	0.427	0.676	1.456	0.506	0.759
CRY2	-0.438	2.042	0.353	0.643	1.274	0.844	0.92
CEBPB	-0.007	1.299	0.378	0.643	0.546	0.15	0.423

Table 4. *Simple interaction format of the non-canonical Wnt module depicting directed edges.*

Edge	From	To
1	PRKCZ	MAPK10
2	FLNA	MAPK10
3	PSME1	PRICKLE1
4	PSME4	PRICKLE1
5	CSK	ITGB5
6	CSK	PPP3CA
7	CSK	MYL6
8	CSK	TPM2
9	CSK	ACTG2
10	CSK	MAPK13
11	CSK	MAPKAPK2
12	CSK	MAPKAPK3
13	RAB11A	AQP2
14	RAB11FIP2	AQP2
15	FURIN	F7
16	ATP6AP2	AGT
17	CTSZ	AGT
18	MME	AGT
19	TRPC1	ITGB5
20	TRPC1	PPP3CA
21	TRPC1	MYL6
22	TRPC1	TPM2
23	TRPC1	ACTG2
24	TBL1X	AGT
25	TBL1X	PPARGC1B
26	TBL1X	NFYC
27	TBL1X	ABCA1
28	TBL1X	CD36
29	TBL1X	CTGF
30	TBL1X	CYP1A1
31	TBL1X	FADS1
32	TBL1X	FHL2
33	TBL1X	GRHL1
34	TBL1X	NPAS2
35	TBL1X	PLIN2
36	UBE2E1	HIST1H2BJ
37	UBE2E1	H3F3B
38	UBE2C	HIST1H2BJ
39	UBE2C	H3F3B
40	UBE2D1	HIST1H2BJ
41	UBE2D1	H3F3B
42	CDK8	CD36
43	MED30	CD36
44	NR3C1	PER2
45	PRKACA	ITGB5
46	PRKACA	PPP3CA
47	PRKACA	MYL6
48	PRKACA	TPM2
49	PRKACA	ACTG2
50	PRKACA	CREB1
51	PRKACA	AQP2
52	PLCB1	ITGB5
53	PLCB1	PPP3CA
54	PLCB1	MYL6
55	PLCB1	TPM2
56	PLCB1	ACTG2
57	PLCB1	BCAR1
58	RASGRP1	WNT5A

Continued...

Edge	From	To
59	RASGRP1	WNT11
60	RASGRP1	FZD5
61	RASGRP1	VAMP2
62	RASGRP1	ITGB5
63	RASGRP1	PPP3CA
64	RASGRP1	MYL6
65	RASGRP1	TPM2
66	RASGRP1	ACTG2
67	RASGRP1	STX1A
68	RASGRP1	GNB1
69	RASGRP1	GNG13
70	RASGRP1	FZD4
71	RASGRP1	BCAR1
72	RAPGEF3	WNT5A
73	RAPGEF3	WNT11
74	RAPGEF3	FZD5
75	RAPGEF3	VAMP2
76	RAPGEF3	ITGB5
77	RAPGEF3	PPP3CA
78	RAPGEF3	MYL6
79	RAPGEF3	TPM2
80	RAPGEF3	ACTG2
81	RAPGEF3	STX1A
82	RAPGEF3	GNB1
83	RAPGEF3	GNG13
84	RAPGEF3	FZD4
85	RAPGEF3	BCAR1
86	CSNK1D	PER2
87	CSNK1D	CRY2
88	SIPA1	ITGB5

Continued...

Edge	From	To
89	SIPA1	PPP3CA
90	SIPA1	MYL6
91	SIPA1	TPM2
92	SIPA1	ACTG2
93	SIPA1	BCAR1
94	ADCY3	ITGB5
95	ADCY3	PPP3CA
96	ADCY3	MYL6
97	ADCY3	TPM2
98	ADCY3	ACTG2
99	ADCY7	ITGB5
100	ADCY7	PPP3CA
101	ADCY7	MYL6
102	ADCY7	TPM2
103	ADCY7	ACTG2
104	ADCY1	ITGB5
105	ADCY1	PPP3CA
106	ADCY1	MYL6
107	ADCY1	TPM2
108	ADCY1	ACTG2
109	ADCY5	ITGB5
110	ADCY5	PPP3CA
111	ADCY5	MYL6
112	ADCY5	TPM2
113	ADCY5	ACTG2
114	PRKCD	ITGB5
115	PRKCD	PPP3CA
116	PRKCD	MYL6
117	PRKCD	TPM2
118	PRKCD	ACTG2

Continued...

Edge	From	To
119	GNAI1	VAMP2
120	GNAI1	ITGB5
121	GNAI1	PPP3CA
122	GNAI1	MYL6
123	GNAI1	TPM2
124	GNAI1	ACTG2
125	GNAI1	STX1A
126	TEAD3	CTGF
127	CTSD	AGT
128	PLCB4	PPP3CA
129	WNT5A	RAC1
130	WNT5A	DVL2
131	WNT5A	CSNK1A1
132	WNT5A	MAPK10
133	WNT5A	DVL1
134	WNT5A	FZD5
135	WNT5A	VAMP2
136	WNT5A	ITGB5
137	WNT5A	PPP3CA
138	WNT5A	MYL6
139	WNT5A	TPM2
140	WNT5A	ACTG2
141	WNT5A	STX1A
142	WNT5A	GNB1
143	WNT5A	GNG13
144	WNT5A	FZD4
145	WNT11	FZD5
146	WNT11	VAMP2
147	WNT11	ITGB5
148	WNT11	PPP3CA

Continued...

Edge	From	To
149	WNT11	MYL6
150	WNT11	TPM2
151	WNT11	ACTG2
152	WNT11	STX1A
153	WNT11	GNB1
154	WNT11	GNG13
155	WNT11	FZD4
156	DVL2	RAC1
157	DVL2	CSNK1A1
158	DVL2	PRICKLE1
159	DVL1	RAC1
160	DVL1	CSNK1A1
161	ROR2	RAC1
162	ROR2	DVL2
163	ROR2	CSNK1A1
164	ROR2	MAPK10
165	ROR2	DVL1
166	FZD5	WNT5A
167	FZD5	WNT11
168	FZD5	DVL2
169	FZD5	DVL1
170	FZD5	VAMP2
171	FZD5	ITGB5
172	FZD5	PPP3CA
173	FZD5	MYL6
174	FZD5	TPM2
175	FZD5	ACTG2
176	FZD5	STX1A
177	FZD5	GNB1
178	FZD5	GNG13

Continued...

Edge	From	To
179	VAMP2	ITGB5
180	VAMP2	PPP3CA
181	VAMP2	MYL6
182	VAMP2	TPM2
183	VAMP2	ACTG2
184	VAMP2	STX1A
185	PPP3CA	MYL6
186	MYL6	ITGB5
187	MYL6	MYL6
188	MYL6	TPM2
189	MYL6	ACTG2
190	STX1A	VAMP2
191	STX1A	ITGB5
192	STX1A	PPP3CA
193	STX1A	MYL6
194	STX1A	TPM2
195	STX1A	ACTG2
196	GNB1	WNT5A
197	GNB1	WNT11
198	GNB1	FZD5
199	GNB1	VAMP2
200	GNB1	ITGB5
201	GNB1	PPP3CA
202	GNB1	MYL6
203	GNB1	TPM2
204	GNB1	ACTG2
205	GNB1	STX1A
206	GNB1	GNG13
207	GNB1	FZD4
208	GNB1	BCAR1

Continued...

Edge	From	To
209	GNG13	WNT5A
210	GNG13	WNT11
211	GNG13	FZD5
212	GNG13	VAMP2
213	GNG13	ITGB5
214	GNG13	PPP3CA
215	GNG13	MYL6
216	GNG13	TPM2
217	GNG13	ACTG2
218	GNG13	STX1A
219	GNG13	GNB1
220	GNG13	FZD4
221	GNG13	BCAR1
222	FZD4	WNT5A
223	FZD4	WNT11
224	FZD4	DVL2
225	FZD4	VAMP2
226	FZD4	ITGB5
227	FZD4	PPP3CA
228	FZD4	MYL6
229	FZD4	TPM2
230	FZD4	ACTG2
231	FZD4	STX1A
232	FZD4	GNB1
233	FZD4	GNG13
234	CREB1	PPARGC1A
235	BCL10	ITGB5
236	BCL10	PPP3CA
237	BCL10	MYL6
238	BCL10	TPM2

Continued...

Edge	From	To
239	BCL10	ACTG2
240	NPAS2	PER2
241	NPAS2	F7
242	PPARGC1A	CD36
243	PPARGC1A	NPAS2
244	HIST1H2BJ	H3F3B
245	H3F3B	HIST1H2BJ
246	SOS1	WNT5A
247	SOS1	WNT11
248	SOS1	FZD5
249	SOS1	VAMP2
250	SOS1	ITGB5
251	SOS1	PPP3CA
252	SOS1	MYL6
253	SOS1	TPM2
254	SOS1	ACTG2
255	SOS1	STX1A
256	SOS1	GNB1
257	SOS1	GNG13
258	SOS1	FZD4
259	RPS6KA1	CREB1
260	RPS6KA1	CEBPB
261	RPS6KA2	CREB1
262	RPS6KA2	CEBPB
263	MAPK1	ITGB5
264	MAPK1	PPP3CA
265	MAPK1	MYL6
266	MAPK1	TPM2
267	MAPK1	ACTG2
268	MAPK1	SOS1
269	MAPK1	RPS6KA1
270	MAPK1	RPS6KA2
271	CEBPB	HIST1H2BJ
272	CEBPB	H3F3B

Continued...

Table 5. *Hub targets measured in cell lines (average logarithm of fold-change) and in patient cohort (hazard ration, p-value, false-discovery rate).*

Hub	Cell lines	Patients		
	logFC	HR	P-val	FDR
FN1	0.66	1.15	0.089	0.319
ESR1	-0.288	0.929	0.142	0.403
COPS5	-0.205	1.294	0.075	0.307
HDAC1	-0.3	1.095	0.576	0.743
MYC	0.543	0.965	0.686	0.764
BMI1	0.446	NA	NA	NA
SUMO3	-0.253	1.203	0.184	0.455
RPA2	-0.296	0.497	0.003	0.032
YWHAG	0.25	NA	NA	NA
EZH2	-0.337	1.107	0.569	0.743
PAN2	-0.282	0.81	0.388	0.743
UBE2I	-0.217	0.901	0.559	0.743
SH3KBP1	0.752	NA	NA	NA
SOX2	-0.976	1.338	0.498	0.743
BARD1	-0.3	0.756	0.095	0.319
HDAC5	-0.456	0.774	0.359	0.738
ATXN1	-0.34	1.473	0.014	0.074
SRRM2	-0.287	1.052	0.585	0.743
ARRB1	0.65	0.892	0.544	0.743
KDM1A	-0.248	1.145	0.585	0.743
LGR4	0.588	1.4	0.013	0.074
SFPQ	-0.214	0.532	0.002	0.032
CD81	0.326	0.929	0.643	0.743
AR	0.528	0.903	0.347	0.738
PRKDC	-0.214	0.916	0.547	0.743
UBE2D1	0.411	1.109	0.628	0.743
POLR2A	-0.292	1.09	0.702	0.764
FYN	1.332	0.957	0.781	0.803
FBXO25	0.264	NA	NA	NA
YBX1	-0.252	1.088	0.563	0.743
FLNA	0.502	0.838	0.181	0.455
UBQLN4	0.193	0.882	0.585	0.743
THRAP3	-0.268	1.032	0.922	0.922
RAC1	-0.194	1.651	0.005	0.038
FHL2	0.998	0.913	0.253	0.584
EPS15	-0.252	0.463	0.001	0.031
E2F1	-0.325	1.944	0.029	0.136
LRPPRC	0.273	1.316	0.105	0.322
PIK3R1	-0.545	0.97	0.742	0.784
AHCYL1	0.228	1.068	0.642	0.743
PLK1	-0.296	2.149	0.005	0.038

References

Abatangelo, L., Maglietta, R., Distaso, A., D'Addabbo, A., Creanza, T. M., Mukherjee, S., and Ancona, N. (2009). Comparative study of gene set enrichment methods. *BMC Bioinformatics*, 10(1):275.

Abba, M. C., Sun, H., Hawkins, K. A., Drake, J. A., Hu, Y., Nunez, M. I., Gaddis, S., Shi, T., Horvath, S., Sahin, A., and Aldaz, C. M. (2007). Breast Cancer Molecular Signatures as Determined by SAGE: Correlation with Lymph Node Status. *Molecular Cancer Research*, 5(9):881–890.

Abdul-Ghani, M., Dufort, D., Stiles, R., De Repentigny, Y., Kothary, R., and Megeney, L. A. (2011). Wnt11 Promotes Cardiomyocyte Development by Caspase-Mediated Suppression of Canonical Wnt Signals. *Molecular and Cellular Biology*, 31(1):163–178.

Aittokallio, T. and Schwikowski, B. (2006). Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics*, 7(3):243–255.

Albert, R. (2005). Scale-free networks in cell biology. *Journal of Cell Science*, 118(21):4947–4957.

Allocco, D. J., Kohane, I. S., and Butte, A. J. (2004). Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics*, 5(1):18.

Amerongen, R. v. and Nusse, R. (2009). Towards an integrated view of Wnt signaling in. *Development*, 136(19):3205–3214.

Anastas, J. N. and Moon, R. T. (2013). WNT signalling pathways as therapeutic targets in cancer. *Nature Reviews Cancer*, 13(1):11–26.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

Barabási, A.-L. and Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68.

Barrett, T. and Edgar, R. (2006). Mining microarray data at ncbi's gene expression omnibus (geo)*. In Bina, M., editor, *Gene Mapping, Discovery, and Expression*, pages 175–190. Springer.

Barry, W. T., Nobel, A. B., and Wright, F. A. (2008). A statistical framework for testing functional categories in microarray data. *The Annals of Applied Statistics*, 2(1):286–315. arXiv: 0803.3881.

Bayerlová, M., Jung, K., Kramer, F., Klemm, F., Bleckmann, A., and Beißbarth, T. (2015a). Comparative study on gene set and pathway topology-based enrichment methods. *BMC bioinformatics*, 16:334.

Bayerlová, M., Klemm, F., Kramer, F., Pukrop, T., Bleckmann, A., and Beißbarth, T. (In press 2015b). Newly constructed network models of different WNT signaling cascades applied to breast cancer expression data. *PLoS ONE*.

Beissbarth, T. (2006). [18] Interpreting Experimental Results Using Gene Ontologies. In Oliver, A. K. a. B., editor, *Methods in Enzymology*, volume 411 of *DNA Microarrays, Part B: Databases and Statistics*, pages 340–352. Academic Press.

Beißbarth, T. and Speed, T. P. (2004). GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465.

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. Series B (Methodological)*, 57(1):289–300.

Bhalla, K., Hwang, B. J., Dewi, R. E., Ou, L., Twaddel, W., Fang, H.-b., Vafai, S. B., Vazquez, F., Puigserver, P., Boros, L., and Girnun, G. D. (2011). PGC1α Promotes Tumor Growth by Inducing Gene Expression Programs Supporting Lipogenesis. *Cancer Research*, 71(21):6888–6898.

Bhola, N. E., Jansen, V. M., Bafna, S., Giltnane, J. M., Balko, J. M., Estrada, M. V., Meszoely, I., Mayer, I., Abramson, V., Ye, F., Sanders, M., Dugger, T. C., Allen, E. V., and Arteaga, C. L. (2015). Kinome-wide Functional Screen Identifies Role of PLK1 in Hormone-Independent, ER-Positive Breast Cancer. *Cancer Research*, 75(2):405–414.

Butte, A. (2002). The use and analysis of microarray data. *Nature Reviews Drug Discovery*, 1(12):951–960.

Cardama, G. A., Comin, M. J., Hornos, L., Gonzalez, N., Defelipe, L., Turjanski, A. G., Alonso, D. F., Gomez, D. E., and Menna, P. L. (2014). Preclinical Development of Novel Rac1-GEF Signaling Inhibitors using a Rational Design Approach in Highly Aggressive Breast Cancer Cell Lines. *Anti-Cancer Agents in Medicinal Chemistry*, 14(6):840–851.

Castillo-Pichardo, L., Humphries-Bickley, T., De La Parra, C., Forestier-Roman, I., Martinez-Ferrer, M., Hernandez, E., Vlaar, C., Ferrer-Acosta, Y., Washington, A. V., Cubano, L. A., Rodriguez-Orengo, J., and Dharmawardhane, S. (2014). The Rac Inhibitor EHop-016 Inhibits Mammary Tumor Growth and Metastasis in a Nude Mouse Model. *Translational Oncology*, 7(5):546–555.

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, Ö., Anwar, N., Schultz, N., Bader, G. D., and Sander, C. (2011). Pathway commons, a web resource for biological pathway data. *Nucleic acids research*, 39(suppl 1):D685–D690.

Chen, E. I., Hewel, J., Krueger, J. S., Tiraby, C., Weber, M. R., Kralli, A., Becker, K., Yates, J. R., and Felding-Habermann, B. (2007). Adaptation of

Energy Metabolism in Breast Cancer Brain Metastases. *Cancer Research*, 67(4):1472–1486.

Choudhary, C. and Mann, M. (2010). Decoding signalling networks by mass spectrometry-based proteomics. *Nature Reviews Molecular Cell Biology*, 11(6):427–439.

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.

Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, 70(6):066111.

Cowling, V. H., D'Cruz, C. M., Chodosh, L. A., and Cole, M. D. (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.

Critchley, D. R. and Gingras, A. R. (2008). Talin at a glance. *Journal of Cell Science*, 121(9):1345–1347.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D'Eustachio, P., and Stein, L. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(suppl 1):D691–D697.

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

Dai, X., Liu, Z., and Zhang, S. (2015). Over-expression of EPS15 is a favorable prognostic factor in breast cancer. *Molecular BioSystems*, 11(11):2978–2985.

De, A. (2011). Wnt/Ca2+ signaling pathway: a brief overview. *Acta Biochimica et Biophysica Sinica*, 43(10):745–756.

Demir, E., Cary, M. P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D'Eustachio, P., Schaefer, C., Luciano, J., et al. (2010). The biopax community standard for pathway data sharing. *Nature biotechnology*, 28(9):935–942.

Dent, R., Trudeau, M., Pritchard, K. I., Hanna, W. M., Kahn, H. K., Sawka, C. A., Lickley, L. A., Rawlinson, E., Sun, P., and Narod, S. A. (2007). Triple-Negative Breast Cancer: Clinical Features and Patterns of Recurrence. *Clinical Cancer Research*, 13(15):4429–4434.

Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., and Müller, T. (2008). Identifying functional modules in protein–protein interaction networks: an integrated exact approach. *Bioinformatics*, 24(13):i223–i231.

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

Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., Georgescu, C., and Romero, R. (2007). A systems biology approach for pathway level analysis. *Genome Research*, 17(10):1537–1545.

Dutta, B., Wallqvist, A., and Reifman, J. (2012). PathNet: a tool for pathway analysis using topological information. *Source Code for Biology and Medicine*, 7(1):10.

D'Eustachio, P. (2013). Pathway Databases: Making Chemical and Biological Sense of the Genomic Data Flood. *Chemistry & Biology*, 20(5):629–635.

Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178.

Ein-Dor, L., Zuk, O., and Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences*, 103(15):5923–5928.

Evangelou, M., Rendon, A., Ouwehand, W. H., Wernisch, L., and Dudbridge, F. (2012). Comparison of Methods for Competitive Tests of Pathway Analysis. *PLoS ONE*, 7(7):e41018.

Fioretti, F. M., Sita-Lumsden, A., Bevan, C. L., and Brooke, G. N. (2014). Revising the role of the androgen receptor in breast cancer. *Journal of Molecular Endocrinology*, 52(3):R257–R265.

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

Fukuda, K.-i. and Takagi, T. (2001). Knowledge representation of signal transduction pathways. *Bioinformatics*, 17(9):829–837.

Gibbons, F. D. and Roth, F. P. (2002). Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation. *Genome Research*, 12(10):1574–1581.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

Goeman, J. J. and Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987.

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690.

Gu, Z., Liu, J., Cao, K., Zhang, J., and Wang, J. (2012). Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes. *BMC Systems Biology*, 6(1):56.

Hagemann, T., Robinson, S. C., Schulz, M., Trümper, L., Balkwill, F. R., and Binder, C. (2004). Enhanced invasiveness of breast cancer cell lines upon co-cultivation with macrophages is due to TNF-α dependent up-regulation of matrix metalloproteases. *Carcinogenesis*, 25(8):1543–1549.

Haibe-Kains, B., Desmedt, C., Loi, S., Culhane, A. C., Bontempi, G., Quackenbush, J., and Sotiriou, C. (2012). A Three-Gene Model to Robustly Identify Breast Cancer Molecular Subtypes. *Journal of the National Cancer Institute*, 104(4):311–325.

Haibe-Kains, B., Schroeder, M., Bontempi, G., Sotiriou, C., and Quackenbush, J. (2011). genefu: Relevant functions for gene expression analysis, especially in breast cancer. *R/Bioconductor. version: Development (2.12)*.

Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika*, 69(3):553–566.

He, J., Sheng, T., Stelter, A. A., Li, C., Zhang, X., Sinha, M., Luxon, B. A., and Xie, J. (2006). Suppressing Wnt Signaling by the Hedgehog Pathway through sFRP-1. *Journal of Biological Chemistry*, 281(47):35598–35602.

Henry, C., Quadir, A., Hawkins, N. J., Jary, E., Llamosas, E., Kumar, D., Daniels, B., Ward, R. L., and Ford, C. E. (2014). Expression of the novel Wnt receptor ROR2 is increased in breast cancer and may regulate both β-catenin dependent and independent Wnt signalling. *Journal of Cancer Research and Clinical Oncology*, 141(2):243–254.

Holland, J. D., Klaus, A., Garratt, A. N., and Birchmeier, W. (2013). Wnt signaling in stem and cancer stem cells. *Current Opinion in Cell Biology*, 25(2):254–264.

Hoshi, H., Sawada, T., Uchida, M., Saito, H., Iijima, H., Toda-Agetsuma, M., Wada, T., Yamazoe, S., Tanaka, H., Kimura, K., Kakehashi, A., Wei, M., Hirakawa, K., and Wanibuchi, H. (2011). Tumor-associated MUC5ac stimulates in vivo tumorigenicity of human pancreatic cancer. *International Journal of Oncology*, 38(3):619–627.

Howe, L. R. and Brown, A. M. C. (2004). Wnt Signaling and Breast Cancer. *Cancer Biology & Therapy*, 3(1):36–41.

Hu, K., Law, J. H., Fotovati, A., and Dunn, S. E. (2012). Small interfering RNA library screen identified polo-like kinase-1 (PLK1) as a potential therapeutic target for breast cancer that uniquely eliminates tumor-initiating cells. *Breast Cancer Research*, 14(1):1–15.

Hu, Z., Mellor, J., Wu, J., Kanehisa, M., Stuart, J. M., and DeLisi, C. (2007). Towards zoomable multidimensional maps of the cell. *Nature Biotechnology*, 25(5):547–554.

Ideker, T. and Krogan, N. J. (2012). Differential network biology. *Molecular Systems Biology*, 8(1):n/a–n/a.

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18(suppl 1):S233–S240.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., Speed, T. P., et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264.

Jaakkola, M. K. and Elo, L. L. (2015). Empirical comparison of structure-based pathway methods. *Briefings in Bioinformatics*, page bbv049.

Jiang, Z. and Gentleman, R. (2007). Extensions to gene set enrichment. *Bioinformatics*, 23(3):306–313.

Kahn, M. (2014). Can we safely target the WNT pathway? *Nature Reviews Drug Discovery*, 13(7):513–532.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32(suppl 1):D277–D280.

Katoh, M. and Katoh, M. (2006). Cross-talk of WNT and FGF signaling pathways at GSK3ß to regulate ß-catenin and SNAIL signaling cascades. *Cancer Biology & Therapy*, 5(9):1059–1064.

Katoh, M. and Katoh, M. (2007). WNT Signaling Pathway and Stem Cell Signaling Network. *Clinical Cancer Research*, 13(14):4042–4045.

Kennecke, H., Yerushalmi, R., Woods, R., Cheang, M. C. U., Voduc, D., Speers, C. H., Nielsen, T. O., and Gelmon, K. (2010). Metastatic Behavior of Breast Cancer Subtypes. *Journal of Clinical Oncology*, 28(20):3271–3277.

Khatri, P., Draghici, S., Ostermeier, G. C., and Krawetz, S. A. (2002). Profiling Gene Expression Using Onto-Express. *Genomics*, 79(2):266–270.

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Comput Biol*, 8(2):e1002375.

Khramtsov, A. I., Khramtsova, G. F., Tretiakova, M., Huo, D., Olopade, O. I., and Goss, K. H. (2010). Wnt/beta-catenin pathway activation is enriched in basal-like breast cancers and predicts poor outcome. *The American Journal of Pathology*, 176(6):2911–2920.

Kirouac, D. C., Saez-Rodriguez, J., Swantek, J., Burke, J. M., Lauffenburger, D. A., and Sorger, P. K. (2012). Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks. *BMC Systems Biology*, 6(1):29.

Kitano, H., Funahashi, A., Matsuoka, Y., and Oda, K. (2005). Using process diagrams for the graphical representation of biological networks. *Nature Biotechnology*, 23(8):961–966.

Klemm, F., Bleckmann, A., Siam, L., Chuang, H. N., Rietkötter, E., Behme, D., Schulz, M., Schaffrinski, M., Schindler, S., Trümper, L., Kramer, F., Beissbarth, T., Stadelmann, C., Binder, C., and Pukrop, T. (2011). β-catenin-independent WNT signaling in basal-like breast cancer and brain metastasis. *Carcinogenesis*, 32(3):434–442.

Kramer, F., Bayerlová, M., Klemm, F., Bleckmann, A., and Beißbarth, T. (2013). rBiopaxParser—an R package to parse, modify and visualize BioPAX data. *Bioinformatics*, 29(4):520–522.

Kühl, M., Sheldahl, L. C., Malbon, C. C., and Moon, R. T. (2000). Ca2+/Calmodulin-dependent Protein Kinase II Is Stimulated by Wnt and Frizzled Homologs and Promotes Ventral Cell Fates in Xenopus. *Journal of Biological Chemistry*, 275(17):12701–12711.

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, 24(5):719–720.

Lavi, O., Dror, G., and Shamir, R. (2012). Network-Induced Classification Kernels for Gene Expression Profile Analysis. *Journal of Computational Biology*, 19(6):694–709.

Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., and Pietenpol, J. A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of Clinical Investigation*, 121(7):2750–2767.

Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323.

Ma, H., Schadt, E. E., Kaplan, L. M., and Zhao, H. (2011). COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method. *Bioinformatics*, 27(9):1290–1298.

Maciejewski, H. (2013). Gene set analysis methods: statistical models and methodological differences. *Briefings in Bioinformatics*, page bbt002.

Massa, M. S., Chiogna, M., and Romualdi, C. (2010). Gene set analysis exploiting the topology of a pathway. *BMC Systems Biology*, 4(1):121.

McDonald, S. L. and Silver, A. (2009). The opposing roles of Wnt-5a in cancer. *British Journal of Cancer*, 101(2):209–214.

Millar, E. K. A., Graham, P. H., O'Toole, S. A., McNeil, C. M., Browne, L., Morey, A. L., Eggleton, S., Beretov, J., Theocharous, C., Capp, A., Nasser, E., Kearsley, J. H., Delaney, G., Papadatos, G., Fox, C., and Sutherland, R. L. (2009). Prediction of Local Recurrence, Distant Metastases, and Death After Breast-Conserving Therapy in Early-Stage Invasive Breast Cancer Using a Five-Biomarker Panel. *Journal of Clinical Oncology*, 27(28):4701–4708.

Miller, J. A., Cai, C., Langfelder, P., Geschwind, D. H., Kurian, S. M., Salomon, D. R., and Horvath, S. (2011). Strategies for aggregating gene expression data: The collapseRows R function. *BMC Bioinformatics*, 12(1):322.

Miller, J. R., Hocking, A. M., Brown, J. D., and Moon, R. T. (1999). Mechanism and function of signal transduction by the Wnt/beta-catenin and Wnt/Ca2+ pathways. *Oncogene*, 18(55):7860–7872.

Mitra, K., Carvunis, A.-R., Ramesh, S. K., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, 14(10):719–732.

Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., Voichita, C., and Draghici, S. (2013). Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in Physiology*, 4.

Moon, R. T., Kohn, A. D., Ferrari, G. V. D., and Kaykas, A. (2004). WNT and β-catenin signalling: diseases and therapies. *Nature Reviews Genetics*, 5(9):691–701.

Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003). Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3):267–273.

Naeem, H., Zimmer, R., Tavakkolkhah, P., and Küffner, R. (2012). Rigorous assessment of gene set enrichment tests. *Bioinformatics*, 28(11):1480–1486.

Neve, R. M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F. L., Fevr, T., Clark, L., Bayani, N., Coppe, J.-P., Tong, F., Speed, T., Spellman, P. T., DeVries, S., Lapuk, A., Wang, N. J., Kuo, W.-L., Stilwell, J. L., Pinkel, D., Albertson, D. G., Waldman, F. M., McCormick, F., Dickson, R. B., Johnson, M. D., Lippman, M., Ethier, S., Gazdar, A., and Gray, J. W. (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell*, 10(6):515–527.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113.

Nishimura, D. (2001). BioCarta. *Biotech Software & Internet Report: The Computer Software Journal for Scient*, 2(3):117–120.

Oishi, I., Suzuki, H., Onishi, N., Takada, R., Kani, S., Ohkawara, B., Koshida, I., Suzuki, K., Yamada, G., Schwabe, G. C., Mundlos, S., Shibuya, H.,

Takada, S., and Minami, Y. (2003). The receptor tyrosine kinase Ror2 is involved in non-canonical Wnt5a/JNK signalling pathway. *Genes to Cells*, 8(7):645–654.

Onitilo, A. A., Engel, J. M., Greenlee, R. T., and Mukesh, B. N. (2009). Breast Cancer Subtypes Based on ER/PR and Her2 Expression: Comparison of Clinicopathologic Features and Survival. *Clinical Medicine & Research*, 7(1-2):4–13.

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J. F., Stijleman, I. J., Palazzo, J., Marron, J. S., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M., and Bernard, P. S. (2009). Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167.

Patil, P., Bachant-Winner, P.-O., Haibe-Kains, B., and Leek, J. T. (2015). Test set bias affects reproducibility of gene signatures. *Bioinformatics*, page btv157.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., et al. (2000). Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752.

Planas-Silva, M. D., Bruggeman, R. D., Grenko, R. T., and Stanley Smith, J. (2006). Role of c-Src and focal adhesion kinase in progression and metastasis of estrogen receptor-positive breast cancer. *Biochemical and Biophysical Research Communications*, 341(1):73–81.

Rakha, E. A., El-Sayed, M. E., Green, A. R., Lee, A. H. S., Robertson, J. F., and Ellis, I. O. (2007). Prognostic markers in triple-negative breast cancer. *Cancer*, 109(1):25–32.

Ritz, A., Tegge, A. N., Kim, H., Poirel, C. L., and Murali, T. M. (2014). Signaling hypergraphs. *Trends in Biotechnology*, 32(7):356–362.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Robinson, M. D. and Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2):321–332.

Ryan, C. J., Cimermančič, P., Szpiech, Z. A., Sali, A., Hernandez, R. D., and Krogan, N. J. (2013). High-resolution network biology: connecting sequence with function. *Nature Reviews Genetics*, 14(12):865–879.

Sadeghi, A. and Fröhlich, H. (2013). Steiner tree methods for optimal sub-network identification: an empirical study. *BMC Bioinformatics*, 14(1):144.

Sales, G., Calura, E., Cavalieri, D., and Romualdi, C. (2012). graphite - a Bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics*, 13(1):20.

Schaefer, C. F. (2004). Pathway databases. *Annals of the New York Academy of Sciences*, 1020:77–91.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K. H. (2009). PID: the Pathway Interaction Database. *Nucleic Acids Research*, 37(suppl 1):D674–D679.

Smid, M., Wang, Y., Zhang, Y., Sieuwerts, A. M., Yu, J., Klijn, J. G. M., Foekens, J. A., and Martens, J. W. M. (2008). Subtypes of Breast Cancer Show Preferential Site of Relapse. *Cancer Research*, 68(9):3108–3114.

Smyth, G. K. (2005). limma: Linear Models for Microarray Data. In Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A., and Dudoit, S., editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Statistics for Biology and Health, pages 397–420. Springer New York.

Staiger, C., Cadot, S., Kooter, R., Dittrich, M., Müller, T., Klau, G. W., and Wessels, L. F. A. (2012). A Critical Evaluation of Network and Pathway-Based Classifiers for Outcome Prediction in Breast Cancer. *PLoS ONE*, 7(4):e34796.

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(Database issue):D535–D539.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., Rijn, M. v. d., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lønning, P. E., and Børresen-Dale, A.-L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874.

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C. M., Lønning, P. E., Brown, P. O., Børresen-Dale, A.-L., and Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences*, 100(14):8418–8423.

Tarca, A. L., Bhatti, G., and Romero, R. (2013). A Comparison of Gene Set Analysis Methods in Terms of Sensitivity, Prioritization and Specificity. *PLoS ONE*, 8(11):e79217.

Tarca, A. L., Draghici, S., Bhatti, G., and Romero, R. (2012). Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*, 13(1):136.

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-s., Kim, C. J., Kusanovic, J. P., and Romero, R. (2009). A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82.

Tarca, A. L., Romero, R., and Draghici, S. (2006). Analysis of microarray experiments of gene expression profiling. *American Journal of Obstetrics and Gynecology*, 195(2):373–388.

Team, R. C. (2012). R: A language and environment for statistical computing.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling survival data: extending the Cox model*. Springer Science & Business Media.

Tripathi, S. and Emmert-Streib, F. (2012). Assessment Method for a Power Analysis to Identify Differentially Expressed Pathways. *PLoS ONE*, 7(5):e37510.

Uray, I. P., Liang, Y., and Hyder, S. M. (2004). Estradiol down-regulates CD36 expression in human breast cancer cells. *Cancer Letters*, 207(1):101–107.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536.

Venet, D., Dumont, J. E., and Detours, V. (2011). Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome. *PLoS Comput Biol*, 7(10):e1002240.

Vidal, M., Cusick, M., and Barabási, A.-L. (2011). Interactome Networks and Human Disease. *Cell*, 144(6):986–998.

Voduc, K. D., Cheang, M. C. U., Tyldesley, S., Gelmon, K., Nielsen, T. O., and Kennecke, H. (2010). Breast Cancer Subtypes and the Risk of Local and Regional Relapse. *Journal of Clinical Oncology*, 28(10):1684–1691.

Vogel, C. and Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, 13(4):227–232.

Walsh, M. D., Clendenning, M., Williamson, E., Pearson, S.-A., Walters, R. J., Nagler, B., Packenas, D., Win, A. K., Hopper, J. L., Jenkins, M. A., Haydon, A. M., Rosty, C., English, D. R., Giles, G. G., McGuckin, M. A., Young, J. P., and Buchanan, D. D. (2013). Expression of MUC2, MUC5ac, MUC5b, and MUC6 mucins in colorectal cancers and their association with the CpG island methylator phenotype. *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc*, 26(12):1642–1656.

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J., Jatkoe, T., Berns, E. M., Atkins, D., and Foekens, J. A. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365(9460):671–679.

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

Watson, P. H., Pon, R. T., and Shiu, R. P. C. (1991). Inhibition of c-myc Expression by Phosphorothioate Antisense Oligonucleotide Identifies a Critical Role for c-myc in the Growth of Human Breast Cancer. *Cancer Research*, 51(15):3996–4000.

Wu, Z., Zhao, X., and Chen, L. (2009). Identifying responsive functional modules from protein-protein interaction network. *Molecules and Cells*, 27(3):271–277.

Yager, J. D. and Davidson, N. E. (2006). Estrogen Carcinogenesis in Breast Cancer. *New England Journal of Medicine*, 354(3):270–282.

Yang, L., Wu, X., Wang, Y., Zhang, K., Wu, J., Yuan, Y.-C., Deng, X., Chen, L., Kim, C. C. H., Lau, S., Somlo, G., and Yen, Y. (2011). FZD7 has a critical role in cell proliferation in triple negative breast cancer. *Oncogene*, 30(43):4437–4446.

Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N., and Liang, H. (2014). Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature Communications*, 5.

Yao, Y.-d., Sun, T.-m., Huang, S.-y., Dou, S., Lin, L., Chen, J.-n., Ruan, J.-b., Mao, C.-q., Yu, F.-y., Zeng, M.-s., Zang, J.-y., Liu, Q., Su, F.-x., Zhang, P., Lieberman, J., Wang, J., and Song, E. (2012). Targeted Delivery of PLK1-siRNA by ScFv Suppresses Her2+ Breast Cancer Growth and Metastasis. *Science Translational Medicine*, 4(130):130ra48–130ra48.

Ye, X., Wang, Y., Rattner, A., and Nathans, J. (2011). Genetic mosaic analysis reveals a major role for frizzled 4 and frizzled 8 in controlling ureteric growth in the developing kidney. *Development (Cambridge, England)*, 138(6):1161–1172.

Yost, C., Torres, M., Miller, J. R., Huang, E., Kimelman, D., and Moon, R. T. (1996). The axis-inducing activity, stability, and subcellular distribution of beta-catenin is regulated in Xenopus embryos by glycogen synthase kinase 3. *Genes & Development*, 10(12):1443–1454.

Yu, H., Kim, P. M., Sprecher, E., Trifonov, V., and Gerstein, M. (2007). The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics. *PLoS Computational Biology*, 3(4).

Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A. T., Sunshine, M., Narasimhan, S., Kane, D. W., Reinhold, W. C., Lababidi, S., Bussey, K. J., Riss, J., Barrett, J. C., and Weinstein, J. N. (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biology*, 4(4):R28.

Zhang, S., Chen, L., Cui, B., Chuang, H.-Y., Yu, J., Wang-Rodriguez, J., Tang, L., Chen, G., Basak, G. W., and Kipps, T. J. (2012). ROR1 Is Expressed in Human Breast Cancer and Associated with Enhanced Tumor-Cell Growth. *PLoS ONE*, 7(3):e31127.

Zhu, Y., Tian, Y., Du, J., Hu, Z., Yang, L., Liu, J., and Gu, L. (2012). Dvl2-Dependent Activation of Daam1 and RhoA Regulates Wnt5a-Induced Breast Cancer Cell Migration. *PLoS ONE*, 7(5):e37823.

Herewith I declare, that I prepared this PhD Thesis on my own and with no other sources and aids than quoted.

Göttingen, November 20, 2015 Michaela Bayerlová