Computational Integrative Analysis of Biological Networks in Cancer

Chengyu Liu

Research Programs Unit,
Genome-Scale Biology,
Faculty of Medicine
University of Helsinki
Finland

Academic dissertation

To be publicly discussed, with the permission of
the Faculty of Medicine of the University of Helsinki,
in the Haartman Institute, Lecture Hall 2, Haartmaninkatu 3, Helsinki,
on 15 September 2017, at 12 o'clock noon.

Helsinki 2017

Supervisor

Sampsa Hautaniemi, DTech, Professor
Research Programs Unit
Genome-Scale Biology
Faculty of Medicine
University of Helsinki
Finland

Reviewers appointed by the Faculty

Petri Auvinen, PhD, Research Director
Institute of Biotechnology
University of Helsinki
Finland

Kimmo Kaski, PhD, Professor
Department of Computer Science
School of Science
Aalto University
Finland

Opponent appointed by the Faculty

Rune Linding, PhD, Professor
Biotech Research and Innovation Centre
University of Copenhagen
Denmark

ISBN 978-951-51-3585-8 (paperback)
ISBN 978-951-51-3586-5 (PDF)
`http://ethesis.helsinki.fi`
Unigrafia Oy
Helsinki 2017

Abstract

Cancer is one of the most lethal diseases. By 2030, deaths caused by cancers are estimated to reach 13 million per year worldwide. Cancer is a collection of related diseases distinguished by uncontrolled cell division that is driven by genomic alterations. Cancer is heterogeneous and shows an extraordinary genomic diversity between patients with transcriptionally and histologically similar cancer subtypes, and even between tumors from the same anatomical position. The heterogeneity poses great challenges in understanding cancer mechanisms and drug resistance; this understanding is critical for precise prognosis and improved treatments.

Emergence of high-throughput technologies, such as microarrays and next-generation sequencing, has motivated the investigation of cancer cells on a genome-wide scale. Over the last decade, an unprecedented amount of high-throughput data has been generated. The challenge is to turn such a vast amount of raw data into clinically valuable information to benefit cancer patients. Single omics data have failed to fully uncover mechanisms behind cancer phenotypes. Accordingly, integrative approaches have been introduced to systematically analyze and interpret multi-omics data, among which network-based integrative approaches have achieved substantial advances in basic biological studies and cancer treatments.

In this thesis, the development and application of network-based integrative methods are included to address challenges in analyzing cancer samples. Two novel methods are introduced to integrate disparate omics data and biological networks at the single-patient level: PerPAS, which takes pathway topology into account and integrates gene expression and clinical data with pathway information; and DERA, which elevates gene expression analysis to the network level and identifies network-based biomarkers that provide functional interpretation. The performance of both methods was demonstrated using biological experiment data, and the results were validated in independent cohorts.

The application part of this thesis focuses on understanding cancer mechanisms and identifying clinical biomarkers in breast cancer and diffuse large B-cell lymphoma using PerPAS, DERA, and an existing method SPIA. Our experimental results provided insights into underlying cancer mechanisms and potential prognostic biomarkers for breast cancer, and identified therapeutic targets for diffuse large B-cell lymphoma. The potential of the therapeutic targets was verified in *in vitro* experiments.

摘要

癌症是一种复杂的疾病，也是现今最致命的疾病之一。 据推算未来二十年后，在世界范围内，每年将有一千三百万人死于癌症。 癌症是异质性疾病，表现出极大的基因组多样性。 取自不同病人但属于相似亚组的基因组样品呈现出显著的差异性, 甚至取自同一个病人同一个位置的基因组样品也是具有差异性。 理解癌症致病机理和发展过程才能更好地提供精确诊断及治疗。

高通量技术的出现激发了系统分析学和计算工具的发展。但是单一平台的数据不足以全面揭示癌症机理，导致理解癌症机理一直是个极大的挑战。 基于网络的整合方法的出现促进了基础生物的研究和病人的诊治。

这篇论文包括两个部分： 整合方法的开发与应用。 在开发新的整合方法方面, 我们研发了新的整合方法来应对整合数据的挑战并回答癌症研究中的问题。 两个新开发的整合方法有: 1) PerPAS, 是一个体化治疗分析工具, 支持单个病人样品的分析, 并且能整合信号通路和基因表达数据。 2) DERA, 是一个整合细胞网络和基因表达数据的工具。它能把基因表达数据的分析提升到网络层面并能进行单个样品的分析。 这两种新型方法的可用性已经在生物数据应用中得以展示，并且用独立数据验证了发现的结果。

整合方法的应用部分集中在全面整合分析mRNA, miRNA, 信号通路数据, 并在弥漫大B细胞淋巴瘤中识别出新的治疗靶点。 在此方法的应用下，我们发现了几个调控重要的临床存活的细胞通路的靶点。并且这些靶点的可靠性已经被实验验证。

Contents

Publications and author's contributions

Publication I **Chengyu Liu**, Rainer Lehtonen, Sampsa Hautaniemi.
PerPAS: Topology-Based Single Sample Pathway Analysis Method.
IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2017, vol.PP, no.99, pp.1-1, doi:10.1109/TCBB.2017.2679745

Publication II **Chengyu Liu**, Riku Louhimo*, Marko Laakso*, Rainer Lehtonen, Sampsa Hautaniemi.
Identification of sample-specific regulations using integrative network level analysis. *BMC Cancer*, 2015, 15:319, doi:10.1186/s12885-015-1265-2

Publication III Suvi-Katri Leivonen*, Katherine Icay*, Kirsi Jäntti, Ilari Siren, **Chengyu Liu**, Amjad Alkodsi, Alejandra Cervera, Maja Ludvigsen, Stephen Jacques Hamilton-Dutoit, Francesco d'Amore, Marja-Liisa Karjalainen-Lindsberg, Jan Delabie, Harald Holte, Rainer Lehtonen, Sampsa Hautaniemi, and Sirpa Leppä.
MicroRNAs regulate key cell survival pathways and mediate chemosensitivity during progression of diffuse large B-cell lymphoma.
Submitted.

* equal contribution

Author's contributions

Publication I First author initiated the novel concept of PerPAS, which integrates gene expression, pathway, and clinical data at the single-patient level. First author independently designed and implemented the algorithm and performed the case study where five different gene expression cohorts and two pathway databases were analyzed and integrated. The analysis included gene expression data processing, PerPAS experiments on breast cancer, and comparison of PerPAS to other methods. First author interpreted the results and wrote the manuscript.

Publication II First author initiated the novel concept of a sample-specific regulation network that is network generated and specific for each cancer patient. First author independently designed and implemented the algorithm and performed the case studies on breast and ovarian cancer gene expression data. The analysis included processing of breast cancer validation datasets and ovarian cancer datasets, DERA experiments on breast and ovarian cancer data, and comparison of DERA to other methods. First author interpreted the results and wrote the manuscript.

Publication III Fifth author used a pathway analysis tool, SPIA, to integrate pathway and gene expression data. The author also analyzed results from pathway analysis and interpreted the results.

Abbreviations

API	Application programming interface
BCR	B-cell receptor
BioPAX	Biological Pathways Exchange
CGCI	Cancer Genome Characterization Initiative
DE	Differentially expressed
DEG	Differentially expressed gene
DER	Differentially expressed regulation
DERA	Differentially Expressed Regulation Analysis
DLBCL	Diffuse Large B-Cell Lymphoma
DNA	Deoxyribonucleic acid
HER2	Human epidermal growth factor receptor 2
mRNA	messenger ribonucleic acid
miRNA	microRNA
ER	Estrogen receptor
HGS-OvCa	High-grade serous ovarian cancer
KEGG	Kyoto Encyclopedia of Genes and Genomes
LLMPP	Lymphoma/Leukemia Molecular Profiling Project
MAPK	Mitogen-activated protein kinase
MRNet	Minimum Redundancy/Maximum Relevance Network
PerPAS	Personalized Pathway Alteration analysiS
PID	Pathway Interaction Database
PLK1	Polo-like kinase 1
PCR	Polymerase chain reaction
PR	Progesterone receptor
PSRN	Patient-specific regulation network
qRT-PCR	Quantitative real-time reverse transcription polymerase chain reaction
RB	Retinoblastoma
RMA	Robust multi-array average
TCGA	The Cancer Genome Atlas
TNBC	Triple-negative breast cancer
WGCNA	Weighted Gene Co-expression Network Analysis
XML	Extensible Markup Language

1 Introduction

We live in an increasingly connected world. Our connections, friends, neighbors, and colleagues indicate who we are, what we do, and how influential we are. Increase of connections in volume challenges advanced use of such information, which requires efficient representation. Network representation is a collection of connections and is useful to visualize and analyze complex systems. Network analysis can provide insights and improve interpretation. Accordingly, applications of network analysis have emerged in various areas, such as mobile communication networks [1] and biological networks [2, 3].

Biological networks consist of numerous molecules, which interact with each other and display highly diverse dynamics. The dynamics of biological networks, which reflect cell conditions and environmental stimulation [4], are controlled and coordinated by multiple levels of information, such as genetics and transcriptomics [5]. Genetic information is known as the blueprint of life [6]. The transcriptome is considered to be the central component in a cell [7], and a biological network is the abstraction of complex logic in cells [8]. Compared to genetic and genomic data, biological network data provide a number of advantages in aggregating molecular events across network neighborhood or genes in the same pathway, thus improving interpretation and comparability, and facilitating multi-omics data integration [8].

Multi-omics data integration is key to understanding biology [9] and has demonstrated its potential in revealing disease mechanisms and identifying prognostic markers and crucial molecules for targeted therapy [10, 11]. Such potential has been driven by technological advancements that efficiently measure tens of thousands of molecules simultaneously. A deluge of molecular data has been produced and been made publicly available to accelerate the understanding of molecular biology, especially molecular cancer biology. Consortia, such as The Cancer Genome Atlas (TCGA) consortium [10], provide molecular and clinical data from tens of thousands of cancer patients. However, such a massive amount of data has posed challenges to data management, interpretation, and integration [12].

Cancer is one of the most lethal diseases, characterized by uncontrolled cellular growth. Though survival of cancer patients has improved due to earlier diagnosis [13], worldwide cancer fatalities were 8.2 million per year in 2012 and are predicted to reach 13 million per year by 2030 [14]. Our understanding of cancer has grown greatly due to the advancements of tools and combined research efforts from multiple fields, such as biology, medicine, mathematics and computer science [7]. However, the heterogeneity of the cancer genome leads to drug resistance and other challenges in cancer treatments.

Many cancer subtypes have been identified [10, 11], enabling personalized treatments [15, 16]. However, these discoveries have not managed to completely stop patients from experiencing cancer progression, relapse, and metastasis. For instance, breast cancer patients belonging to the human epidermal growth factor receptor 2 (HER2) enriched subtype are treated with HER2 inhibitors, whereas few beneficial therapies have been found for patients with the triple-negative breast cancer (TNBC) subtype. TNBC tumors are usually larger in size, higher grade, more aggressive, and have a higher risk of developing distant metastasis than the other breast cancer subtypes [17]. Recent results show that there are substantial differences among cancer samples that belong to similar subtypes [17, 18], which calls for personalizing treatments based on data integration for individual patients.

The goal of this thesis was to study integrative analysis of transcriptomic data, clinical data, and biological networks to understand cancer mechanisms and identify clinical biomarkers in cancer. This thesis consists of two parts: development of integrative analytical methods and their application. The development part aimed to provide improved computational tools that facilitate a deeper understanding of molecular mechanisms in cancer. The goals of the application part were to advance understanding of cancer mechanisms, to identify prognostic markers for predicting cancer progression, and to suggest crucial molecules for targeted therapy. In addition to scientific publications (Publication I-III), unpublished results demonstrated superior performance of our method in integrating multi-omics data.

2 Molecular biology background

The central dogma of molecular biology was first described by Francis Crick in 1956 and later formalized in 1970 [19]. The central dogma of molecular biology states sequential information transfer from genome to proteins (Figure 1). In essence, genetic information contained in deoxyribonucleic acid (DNA) is transcribed into messenger ribonucleic acid (mRNA) and is subsequently translated into proteins.

Most DNA is locally restricted to the cell nucleus. To make genetic information available to the rest of the cell, double-stranded DNA must be transcribed into single-stranded mRNAs and transported to the cytoplasm. In the cytoplasm, proteins are synthesized in the translation process by ribosomes based on sequence information stored in mRNAs. Three nucleotides make up a codon, which determines an amino acid; a sequence of nucleotides determines amino acid sequence of a polypeptide.

The central dogma reflects how information is transferred among different molecules. However, many exceptions to this dogma have been found. For instance, much of the DNA is transcribed into non-coding RNAs such as microRNAs (miRNA) that are about 22 to 26 nucleotides in length. These non-coding RNAs regulate expression of more than 60% of the protein-coding genes in humans [20].

Proteins are functional units in molecular biology and are involved in every biological process. Kinases are one of the most essential proteins regulating almost all signal transduction processes [21]. Kinases are enzymes that catalyze addition of a phosphate group to a specific substrate. Receptors are another important group of proteins that receive and mediate chemical-signals from the extracellular environment into cells or nucleus. Transcription factors are essential proteins [22] that regulate transcription of genes by binding to specific DNA regions. Transcription factors can either activate or inhibit gene transcription, and one transcription factor can regulate multiple genes.

Gene regulation refers to the procedure of transferring information from genome to proteins [23]; this is not always as straightforward as implied in the central dogma. Gene regulation takes place via transcriptional and post-transcriptional regulation. In transcriptional regulation, regulatory proteins such as transcription factors bind to specific DNA sequences (known as transcription binding sites such as promoters and enhancers) to transcribe mRNAs. Transcriptional regulation is considered the most common form of controlling gene expression [23]. Small RNAs such as miRNAs repress gene transcription at the post-transcriptional regulation stage. miRNAs bind to target gene mRNAs to inhibit mRNA expression. Thus, miRNA regulation is an important complement to the central dogma of molecular biology (Figure 1).

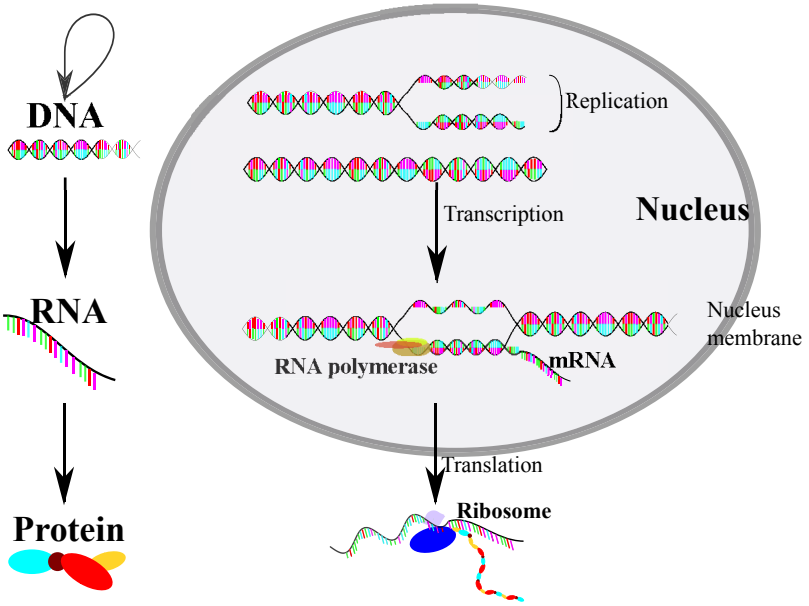

Figure 1: The backbone of molecular biology: **The Central Dogma of Molecular Biology.** The dogma is represented by four major stages. **Replication:** DNA replicates its information in this process. **Transcription:** DNA codes for production of RNAs including messenger RNA (mRNA). In eukaryotic cells, mRNA is processed by splicing, where exons are joined and introns are removed. The mRNA is then delivered from the nucleus to the cytoplasm. **Translation:** mRNA that carries generic information is used as a template to synthesize proteins.

Many cancer studies have been designed at the mRNA level [24, 25, 26, 27]. mRNA measurement is relatively cost efficient, and mRNA quantification is relatively accurate compared to protein expression. Moreover, mRNA expression can be representative of protein expression to some degree [28, 29, 30]. According to the central dogma of molecular biology, mRNA and protein expression should be tightly correlated. While many studies have reported the correlation between mRNA and protein expression [31, 32, 28], studies at the protein level are still necessary, as expression of mRNAs and proteins is not always highly correlated [29, 28]. One reason for low correlation between mRNA and protein expression is that proteins undergo structural changes (known as protein folding) and interact with each other, forming protein complexes. Another reason is the involvement of post-transcriptional regulation, such as miRNA regulation [29]. miRNAs repress protein synthesis by either silencing mRNAs or degrading mRNAs via binding to target gene mRNAs [33].

A pathway is a collection of molecular constituents (including transcription factors,

receptors, and small molecules) and mechanisms through which the molecular constituents are governed, providing various functionalities [34, 8]. Pathways play crucial roles in various physiological and cellular developmental processes [34]. Accordingly, studying pathways is essential to understanding their roles in human diseases, such as cancer [35] and cardiovascular disease [36].

Pathway construction was hindered by the lack of advanced tools and techniques for annotating function of unknown genes and proteins [37]. Nevertheless, the development of cellular and molecular biology experiments and data produced from these experiments are advancing construction of pathways and annotation of elements in the pathways [38]. Experimental observations from the published literature are constantly being mined to improve pathway representations [39, 34, 40]

Another way to construct pathways is to fit mathematical models on biological molecular measurements to infer structures among genes. Many methods have been suggested, such as Minimum Redundancy/Maximum Relevance Networks (MRNET) [41], Weighted Gene Co-expression Network Analysis (WGCNA) [42], and Supervised Inference of Regulatory Networks [43]. Mathematical models provide experimentally testable hypotheses and, in return, biological experiments test these hypotheses and provide experimental data to improve mathematical models in a feedback-loop fashion [44, 45].

3 Biological networks

Network science focuses on studying behaviors of real-world systems using observational data [46]. Networks can be conveniently used to represent complex systems where components are dependent and interact with each other. Accordingly, networks are widely used in many fields, such as technology [1], finance [47], and biology [2, 3]. In biology, networks are applied to data from many levels of measurements resulting in different networks, including protein-protein interaction networks [2], metabolic networks [48], and gene regulation networks [3].

In this chapter, the basics of networks are introduced, followed by a review of biological networks.

3.1 Network basics

In mathematics, networks have been studied under the name of graphs. A network, or a graph *G(V,E)*, is a collection of nodes, or vertices, V, which are connected by a set of links, or edges, $E \subset V \times V$ [49]. In this study, networks and graphs are used interchangeably. A directed network is a network where edges have a direction, while an undirected network is a network where edges do not have orientations. For example, if there is a biological network where vertices represent proteins and edges indicate interactions, then this is an undirected network, as protein A and B interact with each other. In contrast, if the vertices are genes, and there is an edge from gene A to gene B when the product of gene A regulates the expression of gene B, then this network is directed.

A network can have labels and attributes for both vertices and edges, such as names, weights, and types. Vertices and edges can have, in theory, an infinite number of labels and attributes. Attributes can be of numerical or categorical values. Weight, which is normally a numerical attribute, is present in networks called weighted networks. Weights denote different roles in a network. For instance, in a biological network where vertices have a weight attribute that represents the number of neighbors that a vertex connects to, vertices with high weights are much more important than vertices with low weights [48].

Degree of a vertex is the number of edges that the vertex connects to, with self-loops calculated twice [49]. There are three different types of degree (in-, out-, and total). In-degree is the number of in-edge incidents, out-degree is the number of out-edge incidents, and total degree is the sum of in- and out-degrees. Degree is a non-negative value, and an isolated vertex is defined as a vertex with degree zero. A vertex is called a leaf vertex or end vertex when degree is one.

3.1.1 Paths

A path is defined as a number of edges that connect a sequence of vertices in a network [49]. The number of edges in a path can be either finite or infinite. The length of a path is measured by the number of edges in an unweighted network or the sum of edge weights in a weighted network. In a network, it is possible to have many alternative paths from a vertex A to another vertex B. Hence, there are different lengths from vertex A to vertex B. A path, where the number of edges is minimal (in an unweighted network) or the sum of its constituent edge weights is minimized (in a weighted network), is called shortest path given two vertices. Shortest path has many important applications, such as the well-known travelling salesman problem where following question is asked: "A salesman is required to visit once and only once each of N different cities starting from a base city, and returning to this city. What path minimizes the total distance travelled by the salesman?" [50].

3.1.2 Subnetworks

Networks can have a various number of vertices and edges, ranging from zero to a thousand, even a million. When a network is large (e.g., 10,000 vertices), the network can be dissected into small and tractable networks based on functionality or structure. These dissected and small networks are called subnetworks [49]. A subnetwork $S(V_s, E_s)$ of a network $G(V, E)$ is defined as a network where vertices and edges are subsets of vertices and edges of the network G [49]. Subnetworks are particularly useful to study functionality and modularity of networks. Many networks, including social and biological networks, exhibit a high degree of modularity.

3.1.3 Topology

Network topology is the arrangement of vertices and edges in the network. Scale-free and Erdős–Rényi networks are the most common topologies. The term scale-free network, first introduced by Albert-László Barabási and his colleagues, was used to map the topology of the World Wide Web in 1999 [51]. Many networks, such as social and biological networks, have been found to be not random but have features of scale-free networks [51, 52]. A key feature of scale-free networks is the presence of a heavy-tailed degree distribution that follows a power law (Figure 2). Formally, the distribution of scale-free networks ($S(k)$) are defined as below:

$$S(k) = A \cdot k^{\lambda}, \tag{1}$$

where A is a constant value, and k and λ are degree and a degree exponent value, respectively. The value of λ varies depending on the network complexity. For example, the value ranges from two to three in biological networks [52], and $2.1\pm$ 0.1 in a network with over 800 million vertices where a vertex is a document and an edge is a connection pointing to one document from another [51]. The topology of these networks is determined by connectivity of the networks, and hence can be used to effectively locate the most influential molecules in the biological networks and the most informative nodes in the World Wide Web [52, 51].

Scale-free networks have other interesting features such as clustering and hierarchical structure. A direct result of the heavy-tailed degree distribution is indication of a limited number of vertices with degrees that are greatly over mean degree, forming a hierarchical structure. High-degree vertices are often known as hubs and serve specific function in networks, although the functions are dependent mainly on the fields of research.

Scale-free networks show a stunning degree of tolerance against errors. The power law of the degree distribution in a scale-free network implies that the majority of vertices have only one or a few edges, and these vertices with smaller connectivity are targeted with much higher probability when malfunctions and errors occur randomly. Malfunctions and errors in these low-degree vertices do not dramatically change the network structure and have little influence, as the topology of the network almost remains the same. However, robustness against malfunctions and errors comes at a high price, as scale-free networks are lethal to dysfunction of a few vertices (such as hubs) that play key roles in maintaining the network structure [53, 54, 55].

In contrast to scale-free networks, Erdős–Rényi networks, which are named after Paul Erdős and Alfréd Rényi, have a fixed number of vertices with approximately the same number of edges for each vertex [56] (Figure 2). Erdős–Rényi networks are rare in reality and are not covered in this thesis.

3.2 Intracellular networks

Biological networks are used to represent and model chemical reactions in cells, neural connections in nervous systems, and relationships between species in ecosystems [46]. This thesis focuses on the biochemical networks that represent interactions and regulatory mechanisms at the molecular level in biological cells. In particular, this thesis focuses on one of the biochemical networks, gene regulation networks. In this thesis, the biological, biochemical, or intracellular networks refer to gene regulation networks.

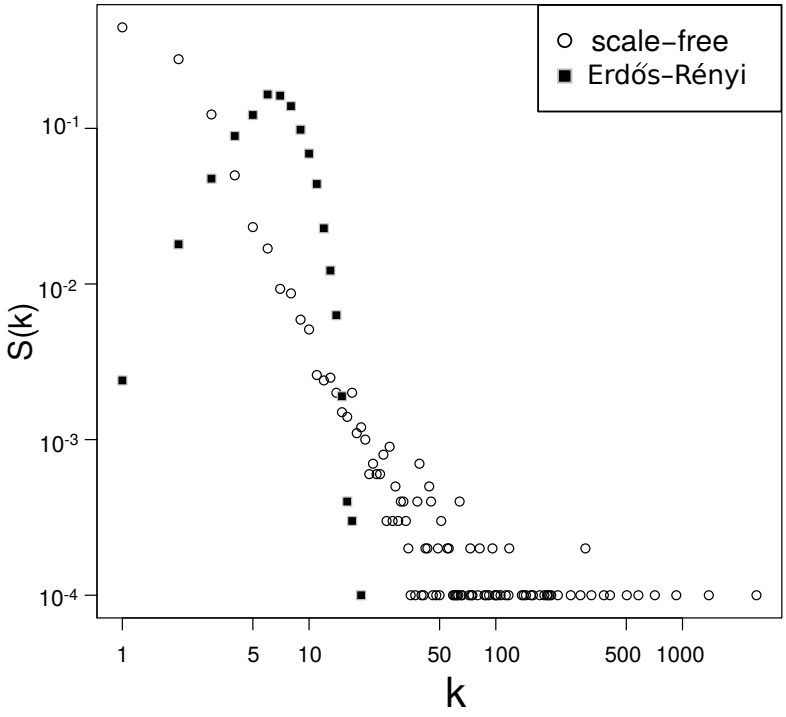

Figure 2: Comparison between degree distributions of scale-free and Erdős-Rényi networks that have an identical number of vertices and edges. Two degree distributions are plotted on a logarithmic scale. The degree distribution of the scale-free networks shows a linear correlation to the degree on the plot, indicating that vertices with lower degree have higher probability. Scale-free networks also have a broad range of degrees, suggesting inhomogeneity. By contrast, the degree distribution of Erdős-Rényi networks peaks at the mean degree and dwindle quickly to both sides, showing that these networks are homogeneous.

3.2.1 Pathways as subnetworks

Pathways can be represented as networks where vertices are genes and small molecules, and edges are regulations among them. Pathways are relatively small compared to intracellular networks, where dynamics of all molecular constituents in a cell are modeled. Pathways are subnetworks of intracellular networks and are assumed to be independent and isolated. Compared to intracellular networks, pathways have advantages in studying functions and interpreting biological systems. On the other hand, pathways normally overlap with each other at a gene or regulation level, and the overlapping genes or regulations often display different functions in

different pathways. Such a phenomenon is referred to as cross-talk. The small scale and isolation of pathways limits understanding from a systems biology perspective; pathways fail to shed light on the whole picture of biological systems.

3.2.2 Gene regulation networks

Gene regulation networks participate in many life processes, including cell differentiation, cell cycle, and apoptosis [3]. Dynamics of gene regulation networks govern gene expression that determines cellular architecture, enzymatic activities, and many other properties through protein expression [23]. Thus, studying patterns of gene regulation networks is crucial to understanding the cellular processes.

Gene regulation networks are a collection of genes and their regulations, which work together to control gene product abundance [3]. In gene regulation networks, genes are represented as vertices and their physical regulations (i.e., gene activations and inhibitions) are represented as edges. Gene regulation networks are directed; a direction from gene A to gene B indicates that gene A is a regulator and controls expression of gene B.

A solid theory of networks provides guidance for exploring mechanisms inside cells from biological networks. Using a network representation form of biology systems, a various number of computational and mathematical approaches (such as graph mining, machine learning, and statistics) can be applied to reveal a variety of insights into biological systems.

3.2.3 Scale-free networks

Studies of topology in biological networks in different species, including humans, have revealed that biological networks are scale-free networks, and the distribution of degree follows power law [51, 52]. High-degree genes (i.e., hub genes) are usually transcription factors or kinase proteins in biological networks. Hub genes are normally the genes that have at least five neighbors or edges [55]. Hub genes play important roles in mediating and controlling signaling flow in biological networks. For example, there are 1895 genes and 5859 regulations in the regulation network generated by merging all the pathways from WikiPathways [39]. The top 15 genes with highest number of neighbors are shown in Figure 3. Out of 15 genes, 12 are either transcription factors or protein kinases. *TP53*, which is one of the most studied genes, has the largest degree, 90 (Figure 3). *TP53* is a transcription factor that has an important role in many anticancer mechanisms such as cell apoptosis [57], genomic stability [58], and inhibition of angiogenesis [59].

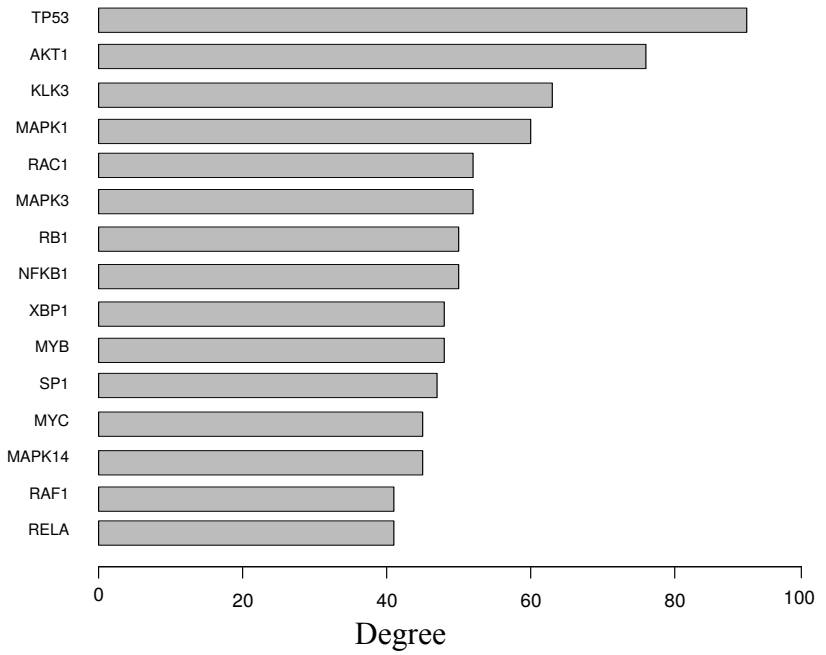

Figure 3: Top 15 genes with highest degree. Degree was calculated from a cellular network merged from WikiPathways. The network only contains gene activations or gene inhibitions from WikiPathways. The network consists of 1895 genes and 5859 regulations between the genes.

Having the property of scale-free networks, biological networks are tolerant to random errors, such as random mutations, but dysfunction of certain genes is lethal. Random mutations are accumulated throughout life. These mutations are equally distributed in the genome, and it is more likely that low-degree genes accumulate many more mutations than high-degree genes. Hence, most people do not show phenotypic effects even though they have several or even hundreds of mutations. However, once mutations occur in key genes, such as hub genes, the damage is severe. For instance, mutations in *TP53* dramatically influence the overall signaling of biological networks, as defects of *TP53* lead to dysregulation of a large number of downstream genes and hinder signaling from one gene to other genes. *TP53* is mutated in about 30% of breast cancer patients [10] and in almost all (96%) high-grade serous ovarian cancer (HGS-OvCa) patients [11].

3.2.4 Bottleneck genes

A recent study has identified another important property of biological networks, which is a bottleneck [61]. A bottleneck measures the amount of signaling

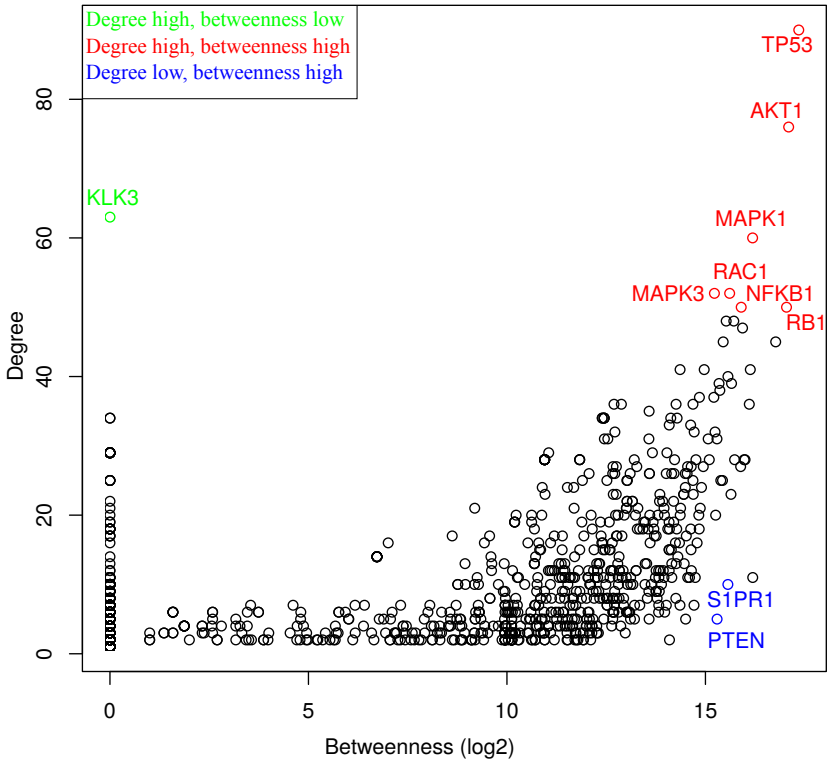

Figure 4: Scatter plot of degree and betweenness. The network only contains gene activations and gene inhibitions from WikiPathways. The network consists of 1895 genes and 5859 regulations between them. In the scatter plot, the X and Y axes represent betweenness and degree, respectively. Betweenness is logarithmic based on two. Degree and betweenness were calculated using an R package igraph [60].

information that goes through a gene. Technically, a bottleneck is evaluated by betweenness centrality that counts the number of shortest paths passing through a gene from all genes to all other genes. The scale of bottlenecks is between zero to hundreds of thousands, while the median value is zero from the gene regulation network generated using WikiPathways database. It has been demonstrated that bottleneck genes play essential roles in controlling and mediating communication information flow from one cluster to another [61]. Bottleneck genes are analogous to bridges and tunnels on a highway map, while hub genes are analogous to roundabouts and highway crossings. Hence, both bottleneck and hub genes are crucial in biological networks.

The degree of bottleneck genes varies, ranging from 2 to 90, as measured from WikiPathways (Figure 4). Here we define a bottleneck as a gene whose betweenness is larger than 2^{10} in the network generated from WikiPathways. Analysis of hub

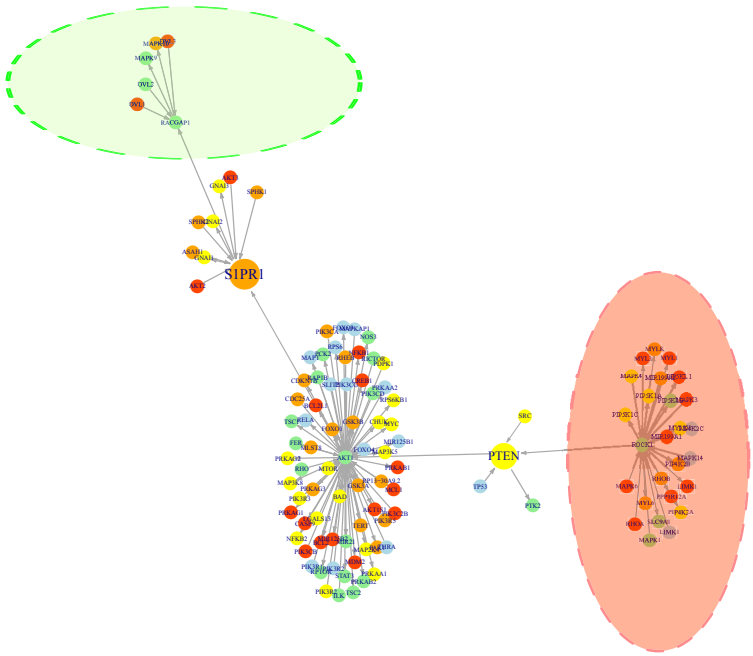

Figure 5: Bottleneck role of *PTEN* and *S1PR1*. This small network is generated from WikiPathways where *PTEN*, *S1PR1*, and their neighbors with distance smaller than two are involved. All signals from the genes in the cluster (red ellipse) must go through *PTEN*, *AKT1*, and *S1PR1* to the genes in another cluster (green ellipse).

and bottleneck genes shows high correlation between them in general (Pearson $r = 0.6$; Figure 4). *TP53* is not only a hub gene but also a bottleneck gene with the largest betweenness and the highest degree. Interestingly, however, high bottleneck genes do not necessarily have high degree, and vice versa (Figure 4). *PTEN* and *S1PR1* have high betweenness but low degree (Figure 4). A subnetwork of the biological network that is related to the *PTEN* and *S1PR1* genes with their neighbors reveals that *PTEN* and *S1PR1* are the main signaling mediators and controllers from one module to another (Figure 5). *S1PR1* is directly regulated by *ATK1*, which is activated by *PTEN*. Thus, all signaling from the module (red cluster in Figure 5) to another (blue cluster in Figure 5) must go through *PTEN, AKT1*, and *S1PR1*. Accordingly, any malfunction in *PTEN, AKT1*, or *S1PR1* completely destroys the signaling from one module to another.

4 Cancer

Cancer is a complex disease characterized by uncontrolled growth and spread of abnormal cells [62]. It is one of the most lethal diseases, and cancer deaths are predicted to rise from an estimated 8.2 million to 13 million per year worldwide by 2030. Cancer is well recognized as a disease of aging. Estimated tumorigenesis occurs at around the 20 years of age and cancer detection at around age 50 [63].

Cancer is partially caused by lifestyle and environmental factors. Unhealthy lifestyles, such as smoking and heavy alcohol consumption, increase the risk of developing cancer [64, 65]. For example, tobacco smokers have a 20 times greater risk of developing lung cancer than non-smokers and have an increased risk of developing many other tumor types as well [66]. Increased exposure to carcinogenic agents present in the occupational and general environment results in an elevated risk of developing cancer. Air pollution, mainly caused by smoke from coal consumption, contributes to a 36-40 times higher lung cancer risk than less-polluted air [67]. Accordingly, the World Health Organization now classifies smoke from coal consumption as a cancer-causing agent.

In this chapter, genomic alterations in cancer are introduced, followed by a discussion of dysregulation of biological networks and integrative approaches in cancer. Finally, cancer heterogeneity is examined.

4.1 Genomic alterations

Cancer is a genomic disease. While an estimated 5% to 10% of all cancers are directly inherited from parents [68, 69], the majority of cancers happen sporadically. Non-hereditary cancers are the focus of this thesis. Non-hereditary cancers arise from accumulated genome instability resulting from random genomic changes [70]. Genomic alterations consist of genetic changes, such as mutations, DNA copy-number alterations, gene expression changes, and epigenetic changes, including histone modifications and DNA methylation.

The hallmarks of cancer [71] are largely driven by genetic and epigenetic alterations [72, 73] through the central dogma of molecular biology. Genetic changes in cancer, such as aberrant expression of oncogenes or tumor-suppressor genes, disturb the protein expression leading to severe consequences. Since DNA copy-numbers are tightly linked to mRNA expression, alterations in DNA copy-numbers change gene expression located in the same DNA regions [74]. DNA point mutations also change mRNA expression by affecting the binding sites of transcription factors [75, 76] or miRNAs [77, 78]. Both DNA copy-number alterations and point mutations may

lead to expression changes of the corresponding proteins. However, genetic changes do not always alter gene expression but may modify protein functions via effects on protein folding and stability [79, 80].

In addition to genetic alterations, epigenetic changes also disturb protein expression in a similar manner by activating cancer genes or inactivating tumor-suppressor genes. Methylation, a type of epigenetic change, plays an important role in cancer through silencing transcription of critical growth regulators (such as tumor-suppressor genes [81]), which subsequently promotes carcinogenesis [82]. Additionally, histone modifications (another type of epigenetic marker) are highly linked to DNA methylation changes and control of gene activity in cancer [83, 84, 85].

Changes in proteins and their expression through either genetic or epigenetic changes affect protein-protein interactions and gene regulation, which eventually impacts the dynamics of biological networks. Furthermore, dysregulation of biological networks results in disruption of fundamental biological processes such as cell death, proliferation, differentiation, and migration [86, 87]. Hence, cancer can be considered a disease of alterations on the biological network level, instead of a single-gene disease [4]. Genetic changes (especially transcriptome changes) and their effects on biological networks are the focus of this thesis.

4.2 Dysregulation of biological networks in cancer

Many forms of cancer exist, and global gene expression changes have been observed in cancer cells compared to normal cells. However, a relatively small number of fundamental alterations are shared by most tumor samples [71]. Mapping genomic alterations onto the pathway level has revealed that cancer samples share more alterations on the network level [88].

Many common pathways are altered in cancer. The cell cycle is one of the most commonly altered pathways in various cancers, including breast cancer [10, 89], ovarian cancer [11, 90] and diffuse large B-cell lymphoma (DLBCL) [91, 92]. The cell cycle governs cell division and DNA replication by tightly regulating cell cycle checkpoints. Hence, dysregulation of the cell cycle impacts cell survival and leads to uncontrolled cell replication, which is one of the major cancer hallmarks [71, 93]. Functional loss of the retinoblastoma gene (*RB*) occurs in many cancers, and disabling pathways related to *RB* is essential for cancer formation due to the tumor-suppressor role of *RB* [94, 95, 96]. Integrative analysis of ovarian cancer by TCGA has shown that the RB pathway is deregulated in 67% of the tumor samples [11]. Mitogen-activated protein kinase (MAPK) pathways, which interconnect extracellular signals, are an evolutionarily conserved mechanism that governs

essential biological processes such as cell growth, proliferation, migration, and apoptosis [97].

4.3 Systematic integrative approaches in cancer

A divide and conquer approach, also known as reductionism, divides complex biological systems into smaller and more manageable constituent parts. Such an approach has been successful for the past 40 years to study the chemical basis and functionality of individual genes or proteins [98]. However, biological systems are complicated and have emergent properties that cannot always be seen on an individual molecular constituent level. In particular, cancer phenotypes cannot be explained by individual molecular constituents [71, 99, 100].

Systems biology was introduced about two decades ago to study holistic and composite properties of biological systems that were undetectable by reductionism, which cannot address the the whole picture of the system [101, 99]. Instead of evaluating a single constituent, systems biology approaches offer simultaneous assessment of many factors of the dynamic system across different time points and contexts [101]. These approaches are becoming a complement to the reductionist approaches.

High-throughput technologies with unprecedented resolution and speed have been developed, and a large number of public high-throughput datasets have been generated since their advent. TCGA is a publicly available repository storing molecular and clinical data of 11,000 patients from 33 different cancer types. In addition to TCGA, the Gene Expression Omnibus (GEO) is another public database that contains data from more than one million samples profiled using mostly microarray and partially next-generation sequencing technologies [102].

Rich genome-scale multi-omics datasets have provided great opportunities to study cancer and have motivated the development of systematic approaches to analyze and integrate the data. Important applications of integrating multi-omics data in cancer research are identification of prognostic biomarkers and cancer subtypes and understanding of cancer mechanisms. Prognostic biomarkers can be used to predict patient survival, cancer subtype identification can provide improved treatments for patients, and understanding of cancer mechanisms can improve interpretation of cancer phenotypes [103].

Many successful applications of integrative analysis on multi-omics data have been reported. One is the identification of subtypes in many cancers. The TCGA consortium has identified subtypes with clinical association (e.g., patient survival) for several cancers, including breast cancer [10] and ovarian cancer [11]. These

subtypes were identified by integrating genetic, transcriptomic, proteomic, and pathway data. Another successful and comprehensive analysis is the identification of breast cancer subtypes by the METABRIC group [104]. Ten novel subgroups were discovered using METABRIC data, where 2,000 breast cancer samples were profiled at both the genomic and transcriptomic levels [104].

Computational integrative approaches help uncover cancer driver genes that are partially or completely responsible for cancer phenotypes [105]. Using a computational framework to detect alterations that promote cancer progression by integrating copy-number and gene expression data, Uri David Akavia and colleagues identified two novel driver genes in melanoma [106]. PARADIGM is a computational tool that determines patient-specific pathway activity by incorporating many types of omics data [104]. PARADIGM outputs pathway-level activity for each patient using probabilistic inference, and its utility has been demonstrated by identifying the clinically relevant subtypes from the glioblastoma multiforme data. iPAS [107] and Pathifier [108] are mathematic integrative approaches that transform gene-level information to biological network-level information. Both methods analyze cancer samples at single-patient resolution, providing biological interpretation on the network level for each patient. Moreover, these two methods have not only revealed clustering with patient overall survival association in glioblastoma multiforme, lung and colorectal cancers, but also provided biological interpretation.

Clinical data play an important role in data integration. The importance of this is shown via building survival association models [109, 110]. Integrating clinical data to build survival models is one of the most widely used approaches. In survival models, survival time and events are correlated with biomarkers (e.g., genes, proteins, pathways). It has been shown that survival models with single molecular data provide little improvements in predicting patient survival, whereas survival models using integrative approaches have much better predictive power [111].

4.4 Heterogeneity in various cancers

Biological variations are any differences between species, individuals, organs, and cells. Some biological variations are visible, such as phenotypic variations including eye color and height. However, variations such as genotypic variations are deeply hidden within the nucleus and are almost invisible as phenotypes. While biological variations have created the diversity of biology and enriched our world, they also increase challenges in health care, especially in cancer treatments. Such variation diversity in cancer is known as cancer heterogeneity. Many cancers encompass a number of histological and genomic subtypes.

More detailed heterogeneity in various cancers will be discussed here: breast cancer, ovarian cancer, and DLBCL. Breast cancer data were used in Publication I and Publication II. Ovarian cancer and DLBCL data were used in Publication II and Publication III, respectively.

Breast cancer is the most common cancer worldwide in females [112]. Breast cancer is an epithelial cancer that develops from cells lining milk ducts. Heterogeneity in breast cancer has been found in both histological and transcriptional profiles and has been known for a long time. Four subtypes, namely TNBC, HER2+, luminal 1, and luminal 2, have been identified using immunohistochemistry based on the expression of estrogen receptor (ER), progesterone receptor (PR), and HER2 [13]. Five subtypes have been stratified using high-throughput gene expression data, namely basal epithelial-like (or basal-like), HER2-enriched, normal breast-like, luminal A, and luminal B groups [113]. There are substantial overlaps between the TNBC and basal epithelial-like subtypes [114, 115, 116]. Subtyping can provide improved and personalized treatments for patients from different subtypes. For example, adjuvant endocrine therapy is used to treat ER-positive patients and leads to a significant improvement in patient overall survival rate and reduction in relapse [117].

TNBC is characterized by low or missing expression of ER, PR, and HER2. TNBC is the most aggressive and invasive breast cancer subtype [17]. There are few beneficial treatments for patients belonging to the TNBC subtype, as patients with the TNBC subtype lack ER and PR expression as targets. Recent studies show that the TNBC subtype can be further divided into six subgroups with different survival associations [17, 118], which further increases the challenge of treating TNBC patients.

Ovarian cancer is an epithelial cancer and is the fifth most lethal cancer in the United States [112]. The estimated number of deaths per year caused by ovarian cancer in United States is 14,180 [112]. HGS-OvCa is the most common and aggressive ovarian cancer subtype. The five-year survival rate of the HGS-OvCa subtype is 35% to 40% [110]. The standard therapy for the HGS-OvCa patients is surgery and platinum-taxane combination chemotherapy. However, most patients who undergo such a treatment relapse after 18 months [119].

HGS-OvCa is genetically characterized by ubiquitous mutations and copy-number alterations [120]. The most common mutations occur in *TP53* (96%) [11]. Germline mutations in *BRCA1* or *BRCA2* are observed in more than 15% of the HGS-OvCa patients, and it has been shown that patients with these mutations have better chemotherapy response [121]. TCGA research has identified four subtypes in ovarian cancer [11], whereas Chen and colleagues have identified three subtypes

[110].

DLBCL belongs to the category of hematological malignancies that are the most common lymphomas in adults. The standard treatment for patients with DLBCL is a combination of rituximab with cyclophosphamide, doxorubicin, vincristine, and prednisone [122, 123]. Despite improved diagnosis and overall outcome of DLBCL patients, an estimated 30-40% of patients experience relapse or resistance to the treatments [124]. This is due to the heterogeneity that exists both among and within the lymphoma subtypes [125, 126]. Patients with DLBCL have been mainly classified into two subtypes, germinal center B-like cell (GCB) and activated B-like cell (ABC) [24]. There is a substantial clinical difference between these two subtypes in five-year survival [127]. Patients from the GCB subtype have less cancer progression and have longer survival time than patients from the ABC subtype [128, 129]. *BCL2* is the most frequently activated oncogene in DLBCL [130]. The phosphatidylinositol signaling system, JAK-STAT cascade, B-cell receptor (BCR) signaling, and MAPK signaling are associated with lymphomas [131, 132].

5 Aims of the study

My research focused on developing and applying computational methods for integrating multi-omics cancer data. In particular, this work focused on methods to integrate transcriptomic, pathway, and clinical data. The general aims were to improve interpretation of transcriptomic data, to identify prognostic markers, and to suggest tailored treatments at the single-patient level.

The specific aims of my research were to:

1. Develop a method that quantifies pathway alterations at a single-patient level by taking pathway topology information into account.
2. Develop a method that integrates transcriptomic data and biological network information at a single-patient level.
3. Apply network-based integrative methods to breast cancer, ovarian cancer, and DLBCL, and to identify putative prognostic markers.

6 Materials and methods

In this chapter, I will summarize the biological materials and computational methods used in each of the publications in this thesis. A more detailed description of materials and methods can be found in each publication.

6.1 Data

An overview of datasets used in my research is presented in Table 1, including cancer types and measurement technologies. For the RNA-Seq gene expression data from TCGA, we used gene expression quantification fully processed by TCGA. For other data from microarray and RNA-Seq technologies, we preprocessed the data ourselves using customized pipelines. In addition, we also used data from the GEO repository to validate the findings from the TCGA data.

Transcriptomic data were used in Publication I, II, and III to quantify differential expression of genes between treatment and control samples. The data were used to study pathway alterations in the treatment samples. Transcriptomic data were analyzed in two steps: preprocessing and differential expression calling.

The preprocessing step of gene expression microarray data consists of background correction where background noise is removed, normalization where chip effects biased by raw probe signals are removed, and summarization in which a set of probe intensities are summarized forming expression of genes. Robust multi-array average (RMA) (that has been used as a standard method) was used for the microarray data [134].

The preprocessing step of RNA-Seq analysis consists of quality control, alignment, and quantification. Quality control is an important step; it trims low-quality bases,

Publication	Cancer	Material	Data type
Publication I	BRCA [10, 102]	Primary tumors	Microarray$*$, RNA-Seq
Publication II	BRCA [10, 102], OvCa [11]	Primary tumors	Microarray$*$, RNA-Seq
Publication III	DLBCL [133, 102]	Primary tumors	Microarray, RNA-Seq$*$

Table 1: Overview of datasets used in each publication. DLBCL: diffuse large B-cell lymphoma; BRCA: breast cancer; OvCa: ovarian cancer. Asterisk ($*$) denotes the datasets that we processed ourselves.

removes remaining tags or adapters from sequencing or polymerase chain reaction (PCR), and discards reads whose length is shorter than a certain threshold. Once this has been completed, reads are aligned to a reference transcriptome. Transcript expression is estimated from the alignment reads. Furthermore, the estimated transcript expression is used to quantify gene expression in the quantification step. RNA-Seq data analysis was performed using Anduril framework, where many sequencing-related components are implemented and customized pipelines can be created [135, 136].

Gene expression measures the relative amount of mRNA quantification but does not indicate if a gene is differentially expressed (DE). Accordingly, differentially expressed genes (DEG) need to be identified. In the differential expression calling step, groups of samples are compared to identify DEGs. One widely used statistic is the t-test, which determines whether two groups of samples are significantly different from one other provided that the samples follow a normal distribution. Another commonly used statistic is fold change, which is calculated as the ratio of two values or means of two groups. Fold change describes the amount of quantity changes from one condition to another.

6.2 Pathway analysis methods

Pathway analysis has been widely used and has experienced three generations over the last 15 years: over-representation analysis, functional class scoring approaches, and pathway topology-based methods [137].

Over-representation analysis, also known as gene set enrichment analysis, was given rise by the need to interpret high-throughput microarray data. Over-representation analysis calculates statistics of a fraction of a predefined gene set enriched among a list of DEGs. One of the main limitations is that over-representation analysis methods consider each gene equally, assuming they are independent from each other [137].

Functional class scoring approaches are an improvement of over-representation analysis methods, and overcome several limitations of the over-representation analysis methods. Functional class scoring approaches treat all genes in a pathway unequally by calculating gene-level statistics. These gene-level statistics are summarized onto pathway-level statistics [137]. One of the most widely used gene set enrichment analysis tools is DAVID [138]. In many cases, pathways contain important information beyond simple gene sets of pathways, such as physical interaction information [39, 34, 40], and neither over-representation analysis nor functional class scoring approaches can integrate such information.

Pathway topology-based approaches are becoming popular and were developed to overcome limitations of over-representation analysis and functional class scoring methods. Pathway topology-based methods calculate gene-level and pathway-level statistics similar to functional class scoring approaches. The key difference is that pathway topology-based methods utilize the gene set of a pathway combined with regulation information among them. A typical tool is SPIA, which introduces pathway impact to analyze signaling pathways [139]. SPIA combines statistics (obtained from classical gene set enrichment analysis) and pathway impact (that measures the significance of pathway perturbation under a given condition).

6.3 Pathway databases

Moksiskaan [140] is a public database that stores pathways from different database repositories including KEGG [34], Pathway Commons [141], and WikiPathways [39]. Moksiskaan provides many useful application programming interfaces (APIs) to integrate connectivity information between genes, proteins, pathways, drugs, and other biological entities, resulting in comprehensive networks. Moksiskaan is built under the Anduril framework [135].

WikiPathways is an open and public web platform used to curate, analyze, and visualize biological pathways for scientific research [39]. WikiPathways supports computational analysis of pathways by providing APIs. WikiPathways stores pathways from different species, including humans. There were 299 human pathways in WikiPathways when we analyzed the data in 2013.

The pathway interaction database (PID) is a freely available collection of curated and peer-reviewed pathways [40], which are composed of human molecular signaling, regulatory events, and key cellular processes. PID offers a range of features to facilitate pathway exploration. These include browsing a predefined set of pathways, creating networks centered on a particular cellular process of interest, and querying lists of molecules derived from high-throughput experiments. In addition, users can also download complete database contents in the format of extensible markup language (XML) or Biological Pathways Exchange (BioPAX). When we analyzed the BioPAX file, 458 pathways were saved in PID in 2015.

While pathway analysis does improve interpretation of high-throughput data, current pathway analysis has several limitations: annotation is incomplete and inaccurate, condition- and cell-specific information is missing, and it is unable to model and analyze dynamic responses [137].

6.4 Personalized cancer patient analysis

Personalized analysis is important to understand cancer mechanisms in individual patients and to apply personalized medicine [15, 16]. One of the key steps in personalized analysis is characterizing individual patient profiles. Patient profiles can be characterized on the gene expression, pathway, or network levels. Characteristics on the gene expression level can be obtained directly from high-throughput measurements, while pathway- and network-level characteristics must be summarized from high-throughput measurements by integrating pathway and network information.

A common strategy to quantify the characteristics of a cancer sample is to quantify the difference between a particular cancer sample and control samples, representing the relative activity of the cancer sample compared to the control samples. Control samples normally are tissue samples from the same organ where the cancer first developed. In an ideal case, a matched control sample, which is a normal tissue sample from the same organ of the same patient, can be used to precisely quantify gene-expression changes in the cancer sample. A matched control sample has minimal biological variations compared to other tissue samples from different patients or from the same patient but different organs. In practice, however, it is difficult to obtain matched control samples due to cost issues or it may be simply impossible (in the case of obtaining brain tissues) [111]. In cases where matched control samples are available, gene expression changes can be represented using fold changes between the cancer and the matched control samples.

In cases of missing matched control samples, accumulated control samples should be adopted to quantify expression changes of genes in cancer samples [142, 108]. The activity of genes in a particular cancer sample can be represented by expression changes (fold changes) between the cancer sample and the mean or median of accumulated control samples. Z-score is used to calculate deviation of gene expression in a particular cancer sample from accumulated control samples, as shown below:

$$Z_{ij} = \frac{E_{ij} - \mu_i}{\sigma_i}, \tag{2}$$

where Z_{ij} represents the activity of gene i in a particular patient j, E_{ij} is the expression measurement of gene i in the patient j, and μ_i and σ_i are the mean and standard deviation of expression of gene i in control samples, respectively.

Patient profiles on the pathway and network levels can be summarized from gene expression changes, such as fold changes or Z-scores. Several summarizing methods have been proposed, such as iPAS and Pathifier. Both methods take gene sets from pathways assuming all genes in a pathway are equal. iPAS calculates an arithmetic

mean of fold changes of all genes involved in a pathway. Pathifier derives a principal curve for each pathway from a dataset, where both cancer and control samples are included, and assigns a pathway-specific score to each sample. However, the roles of genes in the pathways vary, and it is important to reflect the different roles in the pathways when characteristics on the pathway and network levels are summarized. Sophisticated methods not only use gene sets but also pathway topology to present their topological roles in the pathways.

6.5 Kaplan-Meier survival analysis

In medical research, survival analysis is a strategy that answers questions such as: what is the proportion of patients who survive over a period of time, and which group of patients survives better when two or more groups are compared [143]. Survival analysis can be used to assess if a variable (such as a gene, a pathway, or a subnetwork) has sufficient power to predict patient outcomes. In the case of two groups, survival analysis evaluates if a group of patients with a variable survive significantly longer or shorter than another group of patients that do not have the variable. Survival time, also known as follow-up time, is the time period from a beginning point to occurrence of an event of interest, for example the period from cancer diagnosis to the event of relapse, metastasis, or death.

The Kaplan-Meier estimate is one of the best methods for survival analysis. The Kaplan-Meier estimate is often used to measure the probability of patients who live for a certain amount of time after diagnosis or treatments [144]. It is a non-parametric statistic. Formally, the Kaplan-Meier survival function at a given time interval t_i is defined in equation 3:

$$S(t_i) = \frac{n_{t_i}}{N}, \tag{3}$$

where t is a random variable denoting survival time of a patient, n_{t_i} is the number of patients living at time interval t_i, and N is total number of patients alive at the beginning. For each time interval t_i, the survival probability is calculated as the number of patients surviving past t_i divided by the number of patients at risk at the beginning (t_0). At time point t_0, all patients are alive. Patients who drop out of a study are considered as "censored". The Kaplan-Meier survival curve is a decreasing step function, and its theoretical limits are $S(0) = 1$ and $S(\infty) = 0$. The overall probability of survival to a time point is computed by applying multiplication of survival probabilities at all time intervals preceding that time (Figure 6).

A log-rank test is often coupled to a Kaplan-Meier estimate to test the null hypothesis, which states that survival estimates in two or more groups are identical.

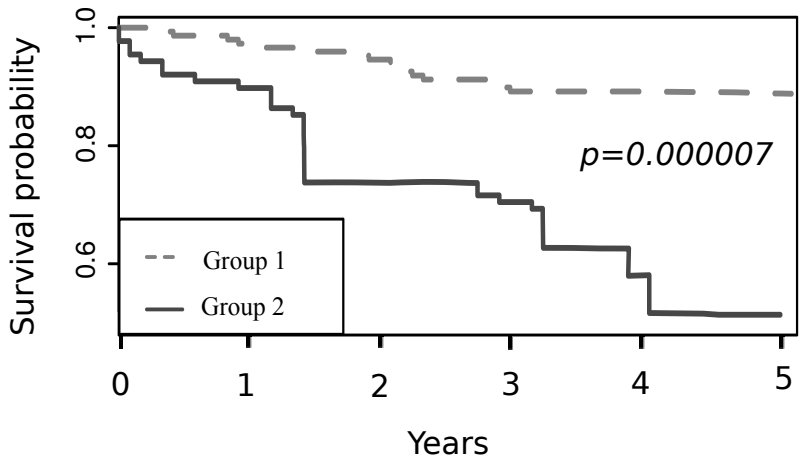

Figure 6: Kaplan-Meier plot. Two groups of samples are compared. Samples from Group 2 survive for shorter periods than samples from Group 1. The X and Y axes represent follow-up time in years and probability of survival, respectively. The survival-associated p-value was calculated using log-rank test.

The log-rank test compares the equality of survival estimates between groups by calculating the expected number of events and the total number of observed events in the groups. With $k \in 2, 3, ...$ patient groups, the test statistic, which follows χ^2 [144], can be used to compute the significance (p-value) of the null hypothesis (Figure 6). In this study, only two-group survival analysis was used.

7 Results

In this chapter, I present the main results on the development of computational integrative analytical methods and their applications in breast cancer, ovarian cancer, and DLBCL. These methods include both an existing tool and novel methods that we have developed to produce the findings in the publications. The main results are the following: 1) PerPAS quantifies pathway activity at a single-patient level to identify pathways that are associated with patient survival, 2) DERA integrates transcriptomic data and biological network information at a single-patient level to identify commonly regulated network modules, and 3) systematic integration of multi-omics data improves data interpretation and identification of potential therapeutic targets.

7.1 Personalized pathway analysis finds putative prognostic markers

Various cancers develop through accumulation of genomic alterations. Study of cancer patients has revealed a great diversity of genetic and transcriptomic profiles, which creates challenges in understanding of cancer mechanisms and treatments. To address these challenges, we developed a novel computational method, PerPAS (**Per**sonalized **P**athway **A**lteration analysi**S**; http://csbi.ltdk.helsinki.fi/pub/czliu/perpas) to interpret large-scale transcriptomic data and to identify altered pathways from individual patients. PerPAS integrates transcriptomic data with clinical and pathway data and quantifies pathway activity for each patient. PerPAS can pinpoint important pathways that are associated with clinical features (such as patient survival) and central nodes in the pathways.

Methodologically, PerPAS first standardizes gene expression data to control samples, indicating deviation of gene expression in a particular cancer sample from the mean of control samples. In cases where control samples are missing from a cohort, gene expression can be standardized to the mean of the cohort. PerPAS then takes advantage of pathway topology information (such as hubness [53, 54, 55] and bottleneckness [61, 55, 145]) to model gene impact on downstream genes. Finally, gene activity is summarized on the pathway level to represent pathway activity of each patient.

To demonstrate the performance of PerPAS and compare it to two existing methods, synthetic and real expression data of the breast cancer patients were used. Synthetic data provide controlled examples to demonstrate the utility of PerPAS and to compare it to other methods, such as iPAS [142] and Pathifier [108] that both

function on the single-patient level. We constructed three small pathways with varying topology specifically to demonstrate the advantages of integrating topology information to the model.

Analysis on synthetic pathways showed that PerPAS assigned various contributions to genes in three pathways based on their topological roles in mediating and controlling signaling. On the other hand, iPAS and Pathifier assigned equal weights to each gene, even though it has been shown that hubness and bottleneckness play important roles in biological networks [53, 54, 55, 61, 145].

The performance of PerPAS was then compared to iPAS and Pathifier using real breast cancer data and was evaluated in terms of its ability to identify pathways that are associated with patient survival. To compare PerPAS, iPAS, and Pathifier, we selected a similar number of significantly altered pathways using different cutoffs resulting in 40 pathways for PerPAS (adjusted $p < 10^{-60}$), 40 for iPAS (adjusted $p < 10^{-60}$), and 43 for Pathifier (adjusted $p < 10^{-140}$)).

PerPAS identified four pathways that were significantly associated with breast cancer patient survival from the TCGA dataset; the association was verified in three or all four independent cohorts. On the other hand, two survival-associated pathways identified by iPAS were validated in one independent cohort. One reason for the low validation rate of iPAS is that an arithmetic mean of gene expression is computed for each pathway, and it is most likely that the averaged gene expression tends to be zero due to the fact that the cancer pathways are composed of overexpressed (positive) and underexpressed (negative) genes. These positive and negative values have more biological meaning than mathematics. Many pathways were thus overlooked by taking the average of gene expression, which neutralized overexpression and underexpression effects. Moreover, due to employing the arithmetic mean of gene expression, iPAS is easily affected by outliers, as all genes are assumed to be equal in the pathways.

Pathifier identified seven survival-associated pathways in the TCGA cohort, however, their associations were not validated due to lack of control samples in the independent cohorts (the main drawback of Pathifier). A requirement for control samples in Pathifier limits its flexibility in many applications. In many cases, it is challenging to obtain control samples. For example, it is rare to have reference brain tissues in glioblastoma multiforme studies.

It was surprising to observe that there was only one overlapping pathway between PerPAS and iPAS, one between iPAS and Pathifier, and none between PerPAS and Pathifier. While taking pathway topology into account is the main reason for the discrepancy between PerPAS and iPAS or Pathifier, it is not the only reason as there was only one overlapping pathway between iPAS and Pathifier. To explore reasons

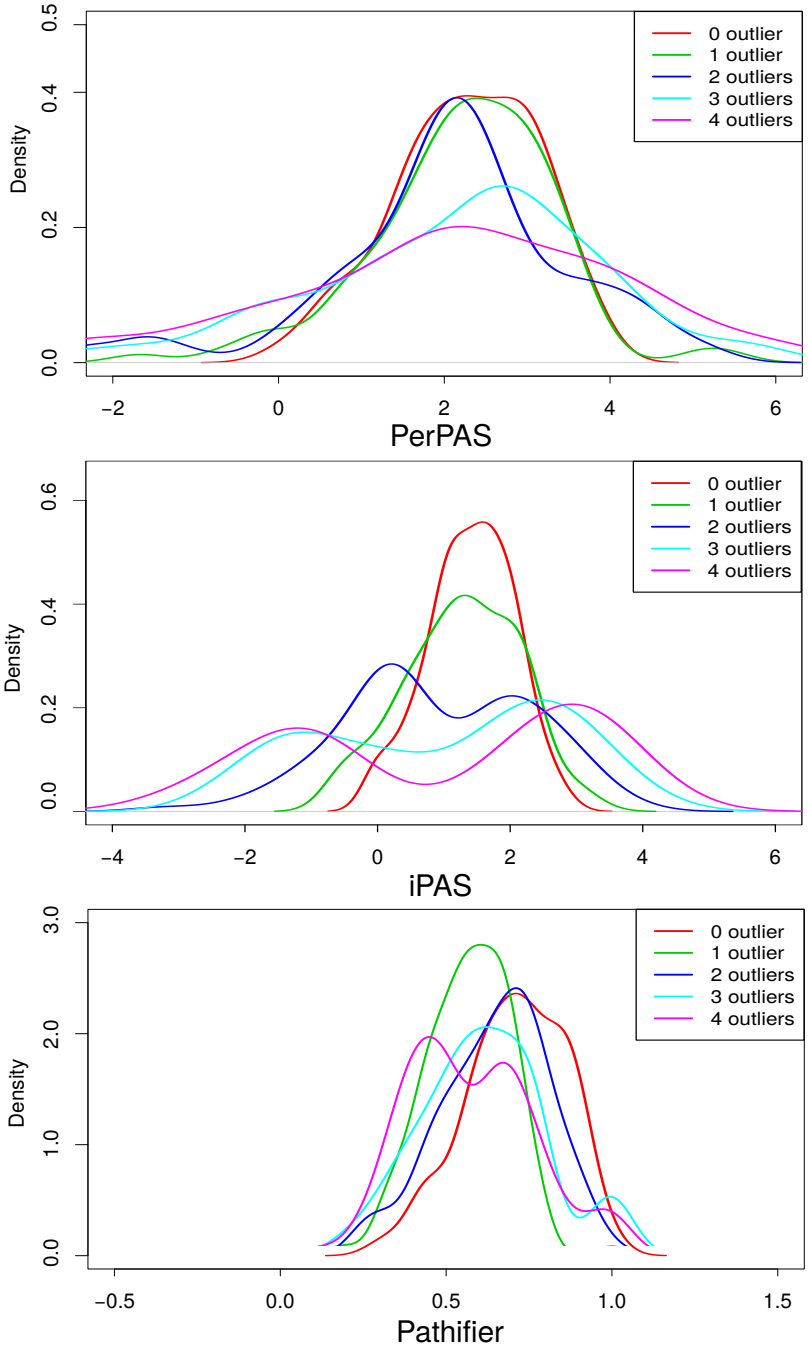

Figure 7: Pathway activity score distributions of a pathway with different numbers of outliers. The X axis denotes pathway activity scores and the Y axis denotes density.

that PerPAS, iPAS and Pathifier produced almost exclusive results, we analyzed the tolerance of the three methods to outliers. We randomly selected 100 treatment and 100 control samples and introduced zero to four outliers to each treatment sample. These outliers were randomly assigned to genes in a pathway that consisted of 41 genes. The pathway activity scores for the particular pathway were calculated using PerPAS, iPAS, and Pathifier.

The density distributions of pathway activity scores by PerPAS did not show clear mean shifting or shape changing until the number of outliers was increased to three (Figure 7). However, the density distribution with one outlier was significantly shifted from that with zero outliers in the Pathifier analysis; density became a bimodal distribution when there were two or more outliers in the iPAS scores. Furthermore, a Kolmogorov-Smirnov test was applied to statistically confirm our observation. The statistical test compared the distribution of pathway activity scores without any outliers to the distributions with different numbers of outliers. The results showed that there were no significant statistical differences until there were up to four outliers in the PerPAS scores (Table 2). The Pathifier results showed significant differences for all comparisons; the iPAS results showed significant differences already with two outliers. These results suggest that outliers have a greater effect on pathway activity scores for iPAS and Pathifier than PerPAS, and tolerance to outliers may be another major reason for the discrepancy between PerPAS and iPAS or Pathifier. The tolerance to outliers may also be the main reason for poor reproducibility in iPAS.

As an example of a PerPAS study, we comprehensively studied the PLK1 signaling events pathway. The PLK1 signaling events pathway was significantly altered in tumor samples compared to control samples. The mean activity score in cancer samples was 3.4 times greater than that in control samples. Our patient survival association study showed that patients with high activity of the PLK1 signaling events pathway exhibited poorer survival than patients who a had lower alteration of this pathway, suggesting its prognostic value in breast cancer. This survival

Methods	One outlier	Two outliers	Three outliers	Four outliers
PerPAS	0.99	0.58	0.09	0.02
iPAS	0.21	2.25E-07	9.57E-06	7.82E-09
Pathifier	1.87E-08	7.80E-02	7.87E-04	4.96E-07

Table 2: Statistical comparison of distributions between pathway activity scores with and without outliers. A Kolmogorov–Smirnov test was used to test the equality of probability distributions. The distribution of pathway activity scores without any outliers was compared to the pathway activity scores with one to four outliers.

association was validated in all four independent cohorts. The prognostic value of the PLK1 signaling events pathway was compared to the *PLK1* gene which has its own prognostic value. The results showed that the PLK1 signaling events pathway had improved prognostic value compared to the *PLK1* gene alone (Table 3).

By further examining the pathway topology, we observed that the *PLK1* gene mediates and controls more than a quarter of signaling, which indicates its central role in this pathway. *PLK1* regulates many known cancer genes, including oncogenes *AURKA* [146] and *ECT2* [147] and tumor-suppressor gene *STAG2* [148]. Furthermore, we found that *PLK1* expression was highly correlated with its downstream genes, suggesting that *PLK1* might directly regulate these downstream genes in the breast cancer patients. The central role of *PLK1* and its high correlation with downstream genes make it a promising and effective therapeutic target. By inhibiting expression of *PLK1*, cancer progression might be repressed. Indeed, *PLK1* is an oncogene [149] and both phase I and II studies of *PLK1*-inhibitory compounds have been conducted [150, 151, 152]. This particular example highlights the potential of PerPAS in providing clinically relevant findings and suggesting putative targets.

To summarize, PerPAS quantifies pathway activity at the single-patients level. PerPAS takes pathway topology into account to score pathway activity. It captures aberrance of pathways compared to control samples and identifies pathways associated with clinical data. We have shown that PerPAS has a much higher validation rate of survival-associated pathways than iPAS or Pathifier. We have further demonstrated that PerPAS can identify both key prognostic pathways and putative therapeutic genes.

Cohorts	*PLK1* gene	PLK1 signaling events pathway
TCGA	0.098	0.004
GSE3494	0.002	0.006
GSE7390	0.003	0.0009
GSE1456	0.0001	0.0000002
GSE4922	0.007	0.009

Table 3: Survival association comparison between the *PLK1* gene and the PLK1 signaling events pathway. The survival-associated p-value was calculated using log-rank test.

7.2 Patient-specific regulation networks enable personalized analysis

DERA (**D**ifferentially **E**xpressed **R**egulation **A**nalysis; http://csbi.ltdk.helsinki.fi/pub/czliu/DERA/) is a novel network approach for integrating biological networks with transcriptomic and clinical data. DERA takes analysis of gene expression data onto the biological network level. Instead of gene sets or individual pathways, where genes across different canonical pathways are interconnected, DERA is able to identify biological network modules that are associated with phenotypes.

DERA first identifies DEGs for each patient and generates patient-specific regulation networks (PSRN) for each patient, by overlaying gene activity status on top of the biological network. Subsequently, these PSRNs are used to identify core differentially expressed regulations (DER) that share similar molecular functions between patients. DERA considers a DER as a core DER if it is identical in at least $F\%$ of all patients. A feature of DERA is a particular fit for analyzing small- or medium-size data that is challenging for many statistical methods where a large number of samples are required. Two types of input data are required in DERA, gene expression data and clinical information (e.g., group or subtype information). The output is a set of core DERs that are phenotype-specific.

We tested DERA with 522 tumors from the TCGA breast cancer microarray cohort, in which 59 control samples were available. The findings were verified in the TCGA breast cancer RNA-Seq cohort, where patients did not overlap with the microarray-profiled patients and the GEO cohort (Table 4). In this case study, we used an F value of 50 in the discovery set; thus DERs must be found and identical in at least 50 % of PSRNs to be categorized as core DERs. A slightly lower F value of 40 % was used because of a smaller size in the GEO cohort. To identify unique regulations that might drive poor prognosis and aggressiveness in TNBC, we compared the core DERs identified from TNBC to those from the other subtypes.

DERA identified 110 core DERs from 55 TNBC samples. Their existence was verified in one or both of the validation sets. Out of 110 core DERs, 22 were validated in both cohorts. These 22 DERs were enriched in cancer-related pathways, such as cell cycle and prostate cancer pathways. 110 core DERs were from different canonical pathways, and the cross-talk issue was overcome. For instance, four pathways (Myometrial Relaxation and Contraction Pathways, Oxidative Stress, Corticotropin-releasing hormone, and TGF-β Signaling Pathway) were connected by *FOS*, suggesting that the connection between canonical pathways would have been unnoticed by just studying individual pathways.

Hierarchical clustering of breast cancer patients from both the TCGA and GEO

	Discovery set	**Validation set**	
	TCGA_array (*n*)	TCGA_seq (*n*)	GEO (*n*)
TNBC	55	56	17
HER2	23	462	26
Luminal 1	219		
Luminal 2	69		
Controls	59	52	7
Total	425	570	50

Table 4: Characteristics of breast cancer datasets. Subtyping is based on immunohistochemistry of ER, PR, and HER2 expression. TNBC: ER-, PR-, and HER2-; HER2: ER-, PR-, and HER2+; Luminal 1: ER+/PR+ and HER2-; Luminal 2: ER+/PR+ and HER2+; Controls: normal breast tissues.

cohorts showed that the genes from 110 core DERs distinguished the TNBC patients from patients with the other subtypes. This result indicates that potential prognostic markers or even drug targets can be derived from these core DERs. Hence, we refined DERs that were unique to TNBC, resulting in 31 DERs. Many of these regulations are related to proliferation, progression, and overall survival in cancer. *FOXA1* is an independent predictor of breast cancer survival and is negatively correlated with tumor size and grade [153]. We found that *XBP1* and its regulator *FOXA1* were underexpressed in the TNBC samples, but overexpressed in the other subtypes. *FOXA1* was strongly and positively correlated with *XBP1* (Pearson $r = 0.83$). We also found that the regulation between *FOXA1* and *XBP1* was associated with breast cancer overall survival; this association was verified in the validation set.

The most interesting regulation was the upregulation of *CCNE1* by *SKP2*. *CCNE1* is a prognostic marker in ER-negative tumors [154], and *SKP2* is an oncogene in breast cancer [155]. We found that *CCNE1* was over-expressed in almost all cancer samples compared to control samples, and the gene expression change in TNBC was even higher compared to the other subtypes. Additionally, *SKP2* was highly expressed only in the TNBC and HER2 subtypes, but not in the luminal 1 or luminal 2 subtypes. These results indicate that higher proliferation and poorer survival in TNBC might be driven by *CCNE1* upregulation, accelerated by further activation of its regulator *SKP2*. Hence, *SKP2* could be a therapeutic target, and *SKP2* inhibition might control cancer cell proliferation and reduce cancer cell survival in TNBC. Indeed, an *SKP2*-inhibitory compound has been identified, and *SKP2* inhibition results in significantly reduced cancer cell proliferation and survival in prostate and lung cancer cells [156].

The utility of our approach was further demonstrated by applying DERA to ovarian cancer gene expression data from TCGA. The application of DERA to ovarian

cancer showed that our method had very high reproducibility. Out of 95 regulations that were identified in the discovery set, 87 (92%) were verified in the validation set. By comparing regulations identified in ovarian cancer and TNBC, we demonstrated that ovarian cancer and TNBC are similar on the molecular level [10].

We also compared DERA with GSEA and SPIA, which are commonly used for pathway analysis. The comparison showed that DERA had a much better validation rate and higher sensitivity in a small group of samples than GSEA and SPIA.

Our main conclusion is that DEGs connect to each other and work synergistically, dysregulating biological networks in breast and ovarian cancers. DERA identified regulations specific for TNBC and suggested a prognostic marker and a therapeutic target for TNBC. Our network-based integrative findings in ovarian cancer gained additional support from the original study on the TCGA data [11]. In conclusion, DERA is capable of identifying solid and potentially clinically valuable regulations, and is comparable with other existing methods.

7.3 Integrative approach interprets transcriptomic data from patients with diffuse large B-cell lymphoma

Integration of multi-omics data is essential to understanding biological processes and mechanisms behind aggressive and progressive cancer phenotypes [157]. To identify potential novel therapeutic targets for relapsed DLBCL patients, we integrated RNA-Seq data from seven paired primary-relapsed samples profiled on the miRNA and mRNA levels. In addition, we used pathway information to identify major pathways associated with cancer progression and survival, and key genes to target.

We found that 13 miRNAs were significantly differentially expressed in the relapsed samples compared to the paired primary samples, suggesting that miRNA expression profiles were quite similar in the primary and relapsed samples. Out of 13 DE-miRNAs, five were excluded from the validation due to undetectably low expression by quantitative real-time reverse transcription polymerase chain reaction (qRT-PCR) or absence of functional primer pairs. Expression of eight DE-miRNAs was tested using qRT-PCR on additional samples, which included 16 matched primary and relapsed DLBCL samples. Five were significantly differentially expressed by qRT-PCR in the primary–relapsed pairs. Many have been reported to suppress tumor growth, invasion, and metastasis, such as *miR-381-3p* [158], *miR-409-3p* [159, 160], and *miR-493-3p* [161, 162]. These reports are consistent with our findings. We found that all five miRNAs were underexpressed in the relapsed samples compared to the paired primary samples, which suggests a lower suppressive effect of these

miRNAs on tumor growth, resulting in tumor growth promotion in the relapsed samples.

To further understand the functional and mechanical roles of the DE-miRNAs, pathway analysis was performed on the target genes of the DE-miRNAs [139]. Given the fact that miRNAs negatively regulate target gene expression [163], we analyzed the correlation between miRNA and transcript expression in seven paired primary-relapsed samples, and subsequently predicted the target genes of 13 DE-miRNAs. We used SPIA for pathway analysis [139].

In total, 1088 transcripts from 787 genes were identified as potential targets of the 13 DE-miRNAs. These 787 target genes were significantly enriched in the pathways related to cancer, such as JAK-STAT cascade, BCR signaling, MAPK signaling, and phosphatidylinositol signaling system. The JAK-STAT cascade has been proposed as a therapeutic target in many tumors [164], such as hematological malignancies [165, 166]. Recent studies also show that the BCR signaling pathway plays an important role in DLBCL pathogenesis [167]. Targeting MAPK signaling inhibits lymphoma cell growth *in vitro* and *in vivo* [168, 169].

To identify key prognostic markers from the enriched pathways, we tested patient overall survival association using the Kaplan-Meier analysis combined with the log-rank test. We performed survival analysis on the Cancer Genome Characterization Initiative (CGCI) ($n = 92$) cohort [24] as the discovery set and the Lymphoma/Leukemia Molecular Profiling Project (LLMPP) cohort ($n = 233$) as the validation set. For survival association, we selected 28 genes targeted by the DE-miRNAs that were enriched in the phosphatidylinositol signaling system, BCR, and MAPK signaling pathways. Out of these 28 genes, six genes were significantly associated with patient overall survival. Two genes, *IMPA1* and *PIP5K1A*, were verified in the validation set.

Five DE-miRNAs were underexpressed in both RNA-Seq and qRT-PCR measurements. Hence, negative regulatory roles of these DE-miRNAs in the relapsed samples might contribute to higher activity of pathways and further result in cancer relapse and drug resistance. We focused on *miR-370-3p* and *miR-409-3p*, which might be regulators in BCR signaling, MAPK signaling pathway and the phosphatidylinositol signaling system. First, we validated four predicted target genes of *miR-370-3p* using qRT-PCR. Elevated expression of *miR-370-3p* induced underexpression of *SYK, PIK3R1, PIK3CD*, and *MAPK1* genes. Next, by studying the effect of *miR-370-3p* and *miR-409-3p* on the proliferation of DLBCL cells, we observed that coexpression of *miR-370-3p* and *miR-409-3p* suppressed cancer cell growth in both ABC and GCB types. Taken together, our clinical and functional study demonstrated that underexpressed miRNAs regulate key cell survival and

drug-resistant pathways in relapsed DLBCL samples via degradation of the mRNAs of the target genes.

In conclusion, integration of miRNA, mRNA expression, and pathway data enables us to identify novel potential therapeutic targets and to interpret the underlying mechanisms. We identified five DE-miRNAs from RNA-Seq data and verified their expression using qRT-PCR from the primary-relapsed pairs. Target genes of the DE-miRNAs were significantly enriched in the pathways related to lymphoma, including phosphatidylinositol signaling system, JAK-STAT cascade, BCR signaling, and MAPK signaling [132, 131]. Two miRNAs, *miR-370-3p* and *miR-409-3p*, were shown to contribute to cancer progression, and activation of *miR-370-3p* combined with *miR-409-3p* significantly suppressed tumor cell growth *in vitro*. We also identified two prognostic markers that were significantly associated with DLBCL patient overall survival.

7.4 Unpublished results: PerPAS simplifies integration and facilitates interpretation of results

Integrative analysis of multi-omics data provides a means of comprehensively understanding mechanisms behind complex phenotypes. Using typical methods, i.e., strategy, statistics, and pathway tool SPIA, integrative analysis has shown potential in identifying prognostic markers and understanding cancer relapse mechanisms (Publication III). In addition, PerPAS identifies altered pathways that are associated with clinical features, and stratifies cancer samples by taking pathway topology information into account (Publication I). To study the role of PerPAS in the integrative analysis approach, PerPAS was applied to DLBCL data. This data contained seven pairs of primary-relapsed samples profiled on both the mRNA and miRNA expression levels. In this study, SPIA was replaced by PerPAS, and integrative analysis using PerPAS focused on the 13 DE-miRNAs identified in Publication III.

We used PerPAS to score pathway activity for each sample using mRNA expression data, and Pearson correlation was adopted to evaluate the anti-correlation role of miRNAs to pathway activity. Out of 443 pathways, 21 pathways that were strongly correlated with the 13 miRNAs (absolute mean of correlation $r > 0.4$) were selected. As expected, many pathways were highly negatively correlated with the 13 miRNAs and related to cancer, such as cell cycle, mTOR signaling, and p75(NTR)-mediated signaling pathways. The strongest negative correlation was observed in the cell cycle pathway. Out of the 13 DE-miRNAs, seven miRNAs were strongly correlated with the cell cycle pathway (Pearson correlation $r < -0.5$); two had even higher correlations (*miR-1226-5p, miR-4664-5p*; Pearson correlation $r < -0.7$). Out of

these seven miRNAs, five were validated using qRT-PCR in the validation set (Publication III).

Unsupervised hierarchical clustering of correlations between the 21 pathways and the 13 miRNAs revealed distinguishable clusters (Figure 8). Interestingly, the group of seven miRNAs was highly anti-correlated with a set of pathways forming a cluster (Figure 8). Some of these pathways (such as cell cycle [10, 89], g-secretase-mediated ERBB4 signaling [170, 171], p75(NTR)-mediated signaling [172, 173], and DCC-mediated attractive signaling pathways [174, 175]) are known to be related to diseases. The cluster where the seven miRNAs were strongly correlated with the six pathways indicates that one miRNA regulates different pathways and one pathway can be regulated by multiple miRNAs. Furthermore, multiple miRNAs may synergistically regulate a set of pathways [20]. Among the seven miRNAs, five were validated using qRT-PCR in the validation cohort (Publication III).

We concluded that *miR-370-3p* combined with *miR-409-3p* suppresses DLBCL cell growth *in vitro* (Publication III). However, the mechanisms behind DLBCL cell growth inhibition via *miR-370-3p* and *miR-409-3p* overexpression is unclear. Integrative analysis of DLBCL using PerPAS instead of SPIA shows that *miR-370-3p* and *miR-409-3p* are clustered together and negatively correlated with a list of pathways. This indicates that complete inhibition of DLBCL cell progression cannot be achieved through activation of a single miRNA but is feasible via overexpression of a combination of miRNAs. Hence, this analysis suggests that overexpression of all seven miRNAs may lead to better inhibition, as alternative miRNA regulations are blocked.

Compared to the integrative analysis used in Publication III, PerPAS provides a higher level of integration by simplifying integration steps and reducing the number of arbitrary thresholds. Instead of identifying correlated genes with DE-miRNAs first and then identifying significant pathways enriched by the correlated genes, miRNA expression was directly correlated with pathway activity. PerPAS uses only one threshold (correlation cutoff), while the typical integrative analysis uses two thresholds (correlation and p-value cutoffs). In addition, PerPAS can identify modules that can facilitate interpretation of complex mechanisms. For example, PerPAS provided evidence for alternative regulations of miRNAs to the pathway cluster. Using typical integrative analysis, however, we failed to decipher the mechanisms where overexpression of both *miR-370-3p* and *miR-409-3p* leads to DLBCL cell growth reduction. Furthermore, PerPAS can provide a rationale based on statistical measurements between pathways and miRNAs, while SPIA analysis provides a list of differentially expressed pathways without providing statistical measurement of the connection between pathways and miRNAs.

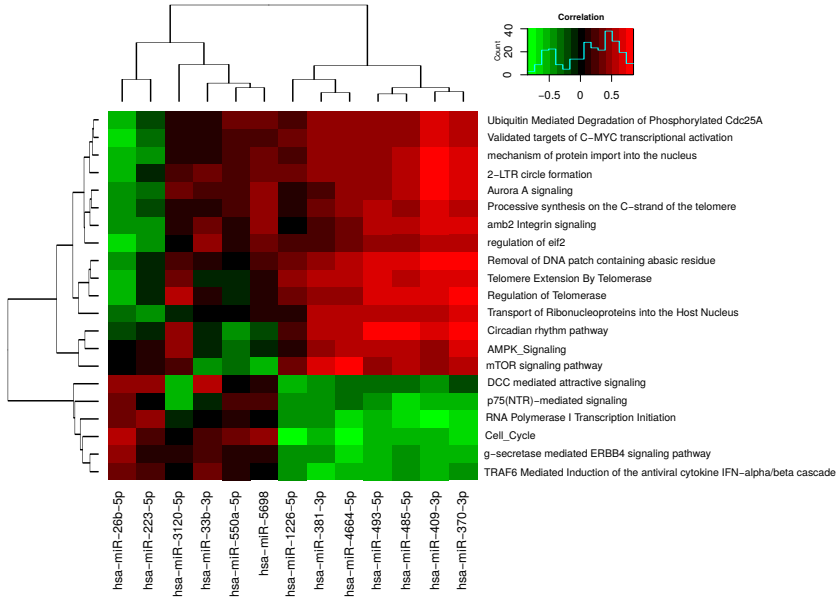

Figure 8: Top 21 pathways correlated with 13 miRNAs. Pathways were selected based on their absolute correlation mean with 13 miRNAs (absolute mean of correlation $r > 0.4$). In the heatmap, rows and columns represent pathways and miRNAs, respectively. Positive and negative correlation is represented as red and green color, respectively.

8 Discussion

Cancer is a heterogeneous disease with characteristics of uncontrolled growth and spread of abnormal cells. Understanding the mechanisms that drive tumorigenesis and drug resistance requires integration of multi-omics data on the single-patient level [176]. Personalized integrative analysis of multi-omics data provides a means to accurately interpret data and to generate solid hypotheses for each patient. This thesis is composed of two parts: the development of novel computational integrative analytical methods to facilitate cancer research and application of these methods to cancer data. In this thesis, I have demonstrated that integrative analysis of biological networks can personalize analysis of heterogeneous cancer samples, identify potential prognostic cancer markers, stratify cancer patients, and improve interpretation of multi-omics data.

Recently, network-based integration approaches have been introduced to substantially advance cancer studies, such as prediction of cancer outcomes [2], identification of potential therapeutic targets [86], and drug development [177]. Thus, more attention has been directed towards development of powerful network-based integrative methods for multi-omics data [5, 9, 178]. Two novel computational network-based integrative methods, PerPAS and DERA, were developed. In this thesis work, these methods exhibit a novel way to integrate gene expression data and biological networks (Publications I and II). PerPAS and DERA take gene expression analysis onto the pathway and network levels, respectively. Both methods aim to analyze multi-omics data on the single-patient level, which has been a challenge in cancer research but is a fundamental step towards personalized medicine [179].

PerPAS was applied to both synthetic and biological gene expression data to demonstrate its utility and performance (Publication I). Application of PerPAS to synthetic data shows that PerPAS is able to integrate pathway structure information to present topological roles of genes in the pathways. The case study of PerPAS on breast cancer gene expression data demonstrated that our method can identify survival-associated biomarkers on the pathway level (which has greater significance compared to molecular biomarkers) and can pinpoint promising targets, such as *PLK1*, in the pathways. When PerPAS was compared to iPAS and Pathifier, the relative performance of these methods was different although all three methods seek to achieve personalized pathway integration [142, 108]. Interestingly, the comparison of the results yielded almost exclusive pathway lists. A detailed comparison among these methods indicated that the higher tolerance of PerPAS to outliers comes from introduction of more information, which illuminates the need for comprehensive data integration.

In addition to integrating miRNA, gene expression, pathway, and clinical data (unpublished results), PerPAS can be used to integrate genetic data to study effects of genetic mutations on the pathway level for each patient and to provide interpretation of genetic mutations. Another potential application of PerPAS is to quantify pathway activity at a single-cell level to study clonal evolution in cancer [180, 181]. In the future, PerPAS could be improved by creating more sophisticated quantifications of gene contribution to pathways. For example, all paths between nodes in a pathway can be used to calculate bottleneck roles, and non-direct downstream genes can be included to quantify hub roles.

Application of DERA on breast and ovarian cancers demonstrated that DERA can identify biomarkers on the network module level, which can subsequently be used to stratify patients and to generate associations with patient survival (Publication II). DERA overcomes the issue of cross-talk between pathways by fusing canonical pathways into a biological network. Instead of individual canonical pathways, DERA identifies network modules that are specific for phenotypes. Identification of network modules provides improved interpretation and precise targets for customized treatments. Indeed, many efforts have been devoted to identify biomarkers on the network-module level. DEAP identifies the most differentially expressed regulatory pattern in the pathways by including pathway structure information [182]. PATHOME detects differentially expressed subpathways by evaluating the significance of differentially expressed paths [183]. Emerging community efforts offer increasing opportunities for improving data integration on the network level.

The DERA method can be improved by identifying connected subnetwork modules instead of individual regulations. An improved DERA can be used to discover novel network modules that may be associated with phenotypes, such as patient survival and mutation status, and to provide new interpretation of results and novel therapeutic treatments for individual patients.

Despite the similar objectives of PerPAS and DERA, these methods are fundamentally different. First, PerPAS integrates gene expression and pathways while DERA integrates gene expression and biological networks. Second, PerPAS identifies biomarkers associated with clinical features on the pathway level, whereas DERA identifies biomarkers on the single-regulation level. Biomarker identification on the pathway level using PerPAS improves interpretation of mechanisms, as the functions of most pathways are known. In contrast, the ability of DERA to identify biomarkers on the regulation level provides more precise targets for biological hypothesis testing. Third, PerPAS summarizes pathway activity for each patient using all genes involved in a pathway and assumes that all genes contribute to pathway activity unequally. DERA, on the other hand, first identifies DEGs for each patient and uses these DEGs to infer patient-specific regulation networks.

A comprehensive integration approach was applied to decipher the relapse mechanisms in DLBCL and to provide predictive and potential therapeutic biomarkers (Publication III). In this study, miRNA and mRNA expression data from paired primary-relapse samples were used, which were valuable as it is challenging to obtain cancer samples from different stages from the same patient. We identified potential therapeutic miRNAs, such as *miR-370-3p* and *miR-409-3p*, which contribute to DLBCL progression by regulating cell survival pathways and affecting chemosensitivity. It has been shown that miRNAs are involved in the development and progression of many cancers, including DLBCL. miRNAs are also used as biomarkers for both prediction and prognosis [184, 185]. Some miRNAs are associated with the DLBCL subtypes and patient survival [186, 187, 26]. However, the potential regulation role of miRNAs in drug resistance and progression in the DLBCL patients is unknown. Our results demonstrated that the miRNAs are clustered and synergistically regulate a set of pathways to drive drug resistance and progression, which would have gone unnoticed without integration of multi-omics data.

All network-based methods presented in this thesis (PerPAS, DERA and SPIA) can integrate multi-omics data to gain insights into complex biology systems. However, they are limited by current biological network databases. It is essential to create high-resolution, accurate, and comprehensive biological network databases with detailed contextual information, such as tissue types and experimental conditions [137, 8]. With advances in high-throughput technologies, high-resolution genomic and proteomic data can be generated. However, the current biological network databases only contain gene-level interactions, and higher resolution such as transcript-level annotation is needed as a gene can be transcribed into many different transcripts, the products of which play roles in biological networks [188]. Regulations, interactions, and pathways are experimentally curated in different tissues under different conditions, hence contextual information is useful to interpret results and to reduce noise. Biological networks are far from complete, but they offer the potential to transform our thinking on biology and personalized cancer treatments [8].

Acknowledgements

This study was carried out in the System Biology Laboratory in the Genome-Scale Biology Program of Research Program Unit, Faculty of Medicine, University of Helsinki during 2010-2016. This study was financially supported by Integrative Life Science Doctoral Graduate Program, Academy of Finland (Center of Excellence in Cancer Genetics Research), Sigrid Juslius foundation, and Finnish Cancer Associations.

First of all, I express my sincerest gratitude to professor Sampsa Hautaniemi. I thank you for introducing me to the world of systems biology and providing sufficient research resources and supportive attitude towards my Master and PhD studies. You have helped me become an independent thinker and scientist. You always encouraged me to come up with my own conclusions and solutions when there were obstacles. Without your insightful guidance, I could not have accomplished my goal.

I thank the members of my thesis committee, professor Tero Aittokallio and docent Petri Auvinen, for their encouragement and advice during the annual meetings.

I wish to thank the reviewers of my thesis, docent Petri Auvinen and professor Kimmo Kaski, for their time and comments to my thesis work.

I want to express my sincere thanks to the friendly and helpful colleagues in the lab: Anna-Maria, Marko, Sirkku, Kristian, Ville, Riku, Ping, Tiia, Javier, Erkka, Lilli, Alejandra, Viljami, Vladimir, Lauri, Rainer, Katherine, Amjad, Chiara, Emilia, Julia, Mikko, and Kaiyang. Special thanks to: Rainer Lehtonen, who gave comments from the point of view of a biologist and helped me improve my biological understanding; Marko Laakso, who was my mentor in all things related to biological networks and the co-worker in many research projects; Ville Rantanen, who taught me about the Linux and helped me solve thousands of related problems; Tiia Pelkonen, who helped me with English language proofreading from my first publication to this thesis. Riku Louhimo and Erkka Valo, whom I played football with after work.

I am thankful to the wet lab collaborators involved in the DLBCL study in this thesis: professor Sirpa Leppä, Suvi-Katri Leivonen, Kirsi Jäntti and Ilari Siren.

Finally, I express my gratitude to my patients, Chengchun and Zaisong, for your immeasurable love and support. You have been selflessly offering the best to me. I would like to thank my brother, Chenghao, for growing up and playing together since my birth. You are my inspirations, and your achievement remains the yardstick for me in my future endeavors. I am grateful to my fiancée Yu for your assistance

and encouragement, and for scaring off the bear that ate my sock. Words cannot express how lucky I am to have you in my life.

Chengyu Liu
Helsinki, 2017

References

Page(s)

[1] Onnela, J.-P, Saramäki, J, Hyvönen, J, et al. (2007) Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences* **104**, 7332–7336. — 1, 6

[2] Taylor, I. W, Linding, R, Warde-Farley, D, et al. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature biotechnology* **27**, 199–204. — 1, 6, 39

[3] Karlebach, G & Shamir, R. (2008) Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology* **9**, 770–780. — 1, 6, 10

[4] Barabási, A.-L, Gulbahce, N, & Loscalzo, J. (2011) Network medicine: a network-based approach to human disease. *Nature Reviews Genetics* **12**, 56–68. — 1, 15

[5] Charitou, T, Bryan, K, & Lynn, D. J. (2016) Using biological networks to integrate, visualize and analyze genomics data. *Genetics, Selection, Evolution* **48**. — 1, 39

[6] McCarthy, N. (2010) Genetics: Back to the blueprint. *Nature Reviews Cancer* **10**, 161–161. — 1

[7] Chen, P. (2016) Ph.D. thesis (University of Helsinki). — 1

[8] Pau, C, Jüri, R, Syed, H, et al. (2015) Pathway and network analysis of cancer genomes. *Nature Methods* **12**, 615–621. — 1, 5, 41

[9] Ebrahim, A, Brunk, E, Tan, J, et al. (2016) Multi-omic data integration enables discovery of hidden biological regularities. *Nature Communications* **7**, 13091. 00005. — 1, 39

[10] Network, T. C. G. A. (2012) Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70. — 1, 2, 11, 15, 16, 21, 34, 37

[11] Network, T. C. G. A. R. (2011) Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615. — 1, 2, 11, 15, 16, 18, 21, 34

[12] Hamid, J. S, Hu, P, Roslin, N. M, et al. (2009) Data Integration in Genetics and Genomics: Methods and Challenges. *Human Genomics and Proteomics : HGP* **2009**. — 1

[13] Onitilo, A. A, Engel, J. M, Greenlee, R. T, et al. (2009) Breast Cancer Subtypes Based on ER/PR and Her2 Expression: Comparison of Clinicopathologic Features and Survival. *Clinical Medicine & Research* **7**, 4–13. — 1, 18

[14] Torre, L. A, Bray, F, Siegel, R. L, et al. (2015) Global cancer statistics, 2012. *CA: A Cancer Journal for Clinicians* **65**, 87–108. — 1

[15] Cho, S.-H, Jeon, J, & Kim, S. I. (2012) Personalized Medicine in Breast Cancer: A Systematic Review. *Journal of Breast Cancer* **15**, 265–272. — 2, 24

[16] Chan, I. S & Ginsburg, G. S. (2011) Personalized medicine: progress and promise. *Annual Review of Genomics and Human Genetics* **12**, 217–244. — 2, 24

[17] Lehmann, B. D, Bauer, J. A, Chen, X, et al. (2011) Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *Journal of Clinical Investigation* **121**, 2750–2767. 2, 18

[18] Ciriello, G, Sinha, R, Hoadley, K. A, et al. (2013) The molecular diversity of Luminal A breast tumors. *Breast Cancer Research and Treatment* **141**, 409–420. 2

[19] Crick, F. (1970) Central Dogma of Molecular Biology. *Nature* **227**, 561–563. 3

[20] Friedman, R. C, Farh, K. K.-H, Burge, C. B, et al. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research* **19**, 92–105. 3, 37

[21] Schenk, P. W & Snaar-Jagalska, B. (1999) Signal perception and transduction: the role of protein kinases. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1449**, 1 – 24. 3

[22] Latchman, D. S. (1997) Transcription factors: an overview. *The International Journal of Biochemistry & Cell Biology* **29**, 1305–1312. 3

[23] Griffiths, A. J, Wessler, S. R, Carroll, S. B, et al. (2010) *An Introduction to Genetic Analysis*. (W. H. Freeman), 10th edition. 3, 10

[24] Alizadeh, A. A, Eisen, M. B, Davis, R. E, et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511. 4, 19, 35

[25] Schena, M, Shalon, D, Davis, R. W, et al. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470. 4

[26] Lim, E. L, Trinh, D. L, Scott, D. W, et al. (2015) Comprehensive miRNA sequence analysis reveals survival differences in diffuse large B-cell lymphoma patients. *Genome Biology* **16**. 4, 41

[27] Ma, X.-J, Salunga, R, Tuggle, J. T, et al. (2003) Gene expression profiles of human breast cancer progression. *Proceedings of the National Academy of Sciences* **100**, 5974–5979. 4

[28] Greenbaum, D, Colangelo, C, Williams, K, et al. (2003) Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biology* **4**, 117. 4

[29] Maier, T, Güell, M, & Serrano, L. (2009) Correlation of mRNA and protein in complex biological samples. *FEBS Letters* **583**, 3966–3973. 4

[30] Zhang, B, Wang, J, Wang, X, et al. (2014) Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382–387. 4

[31] Guo, Y, Xiao, P, Lei, S, et al. (2008) How is mRNA expression predictive for protein expression? A correlation study on human circulating monocytes. *Acta Biochimica Et Biophysica Sinica* **40**, 426–436. 4

[32] Ghaemmaghami, S, Huh, W.-K, Bower, K, et al. (2003) Global analysis of protein expression in yeast. *Nature* **425**, 737–741. 4

[33] Fabian, M. R, Sonenberg, N, & Filipowicz, W. (2010) Regulation of mRNA Translation and Stability by microRNAs. *Annual Review of Biochemistry* **79**, 351–379. 4

[34] Kanehisa, M & Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27–30. 5, 22, 23

[35] Larue, L & Bellacosa, A. (2005) Epithelial–mesenchymal transition in development and cancer: role of phosphatidylinositol 3' kinase/AKT pathways. *Oncogene* **24**, 7443–7454. 5

[36] Hunter, J. J & Chien, K. R. (1999) Signaling pathways for cardiac hypertrophy and failure. *New England Journal of Medicine* **341**, 1276–1283. 5

[37] Chowdhury, S & Sarkar, R. R. (2015) Comparison of human cell signaling pathway databases–evolution, drawbacks and challenges. *Database* **2015**. 5

[38] Hao, T, Ma, H.-W, Zhao, X.-M, et al. (2010) Compartmentalization of the Edinburgh Human Metabolic Network. *BMC Bioinformatics* **11**, 393. 5

[39] Kelder, T, van Iersel, M. P, Hanspers, K, et al. (2012) WikiPathways: building research communities on biological pathways. *Nucleic Acids Research* **40**, D1301–D1307. 5, 10, 22, 23

[40] Schaefer, C. F, Anthony, K, Krupa, S, et al. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Research* **37**, D674–D679. 5, 22, 23

[41] Meyer, P. E, Kontos, K, Lafitte, F, et al. (2007) Information-theoretic inference of large transcriptional regulatory networks. *EURASIP journal on bioinformatics & systems biology* p. 79879. 5

[42] Langfelder, P & Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559. 5

[43] Mordelet, F & Vert, J.-P. (2008) SIRENE: supervised inference of regulatory networks. *Bioinformatics* **24**, i76–82. 5

[44] Nijhout, H. F, Best, J. A, & Reed, M. C. (2015) Using mathematical models to understand metabolism, genes, and disease. *BMC Biology* **13**, 79. 5

[45] Sulaimanov, N, Klose, M, Busch, H, & Boerries, M. (2017) Understanding the mtor signaling pathway via mathematical modeling. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* pp. e1379–n/a. 5

[46] Newman, M. (2010) *Networks: an introduction*. (OUP Oxford). 6, 8

[47] Onnela, J.-P, Kaski, K, & Kertesz, J. (2003) Clustering and information in correlation based financial networks. *The European Physical Journal B* **38**. 6

[48] Jeong, H, Tombor, B, Albert, R, et al. (2000) The large-scale organization of metabolic networks. *Nature* **407**, 651–654. 6

[49] Diestel, R. (2010) *Graph Theory, Fourth Edition*. (Springer). 6, 7

[50] Bellman, R. (1962) Dynamic Programming Treatment of the Travelling Salesman Problem. *J. ACM* **9**, 61–63. 7

[51] Barabási, A.-L & Albert, R. (1999) Emergence of Scaling in Random Networks. *Science* **286**, 509–512. 7, 8, 10

[52] Albert, R & Barabási, A.-L. (2002) Statistical mechanics of complex networks. *Reviews of Modern Physics* **74**, 47–97. 7, 8, 10

[53] Albert, R, Jeong, H, & Barabási, A.-L. (2000) Error and attack tolerance of complex networks : Article : Nature. *Nature* **406**, 378–382. 8, 27, 28

[54] Jeong, H, Mason, S. P, Barabási, A.-L, et al. (2001) Lethality and centrality in protein networks. *Nature* **411**, 41–42. 8, 27, 28

[55] Dietz, K.-J, Jacquot, J.-P, & Harris, G. (2010) Hubs and bottlenecks in plant molecular signalling networks. *New Phytologist* **188**, 919–938. 8, 10, 27, 28

[56] Erdös, P & Rényi, A. (1959) On random graphs, I. *Publicationes Mathematicae (Debrecen)* **6**, 290–297. 8

[57] Vazquez, A, Bond, E. E, Levine, A. J, et al. (2008) The genetics of the p53 pathway, apoptosis and cancer therapy. *Nature Reviews Drug Discovery* **7**, 979–987. 10

[58] Negrini, S, Gorgoulis, V. G, & Halazonetis, T. D. (2010) Genomic instability — an evolving hallmark of cancer. *Nature Reviews Molecular Cell Biology* **11**, 220–228. 10

[59] Teodoro, J. G, Parker, A. E, Zhu, X, et al. (2006) p53-mediated inhibition of angiogenesis through up-regulation of a collagen prolyl hydroxylase. *Science* **313**, 968–971. 10

[60] Csardi, G & Nepusz, T. (2006) The igraph Software Package for Complex Network Research. *InterJournal* p. 1695. 12

[61] Yu, H, Kim, P. M, Sprecher, E, et al. (2007) The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics. *PLoS Computational Biology* **3**, e59. 11, 12, 27, 28

[62] Weinberg, R. (2013) *The Biology of Cancer, Second Edition*. (Garland Science). 14

[63] Anand, P, Sundaram, C, Jhurani, S, et al. (2008) Curcumin and cancer: an "old-age" disease with an "age-old" solution. *Cancer Letters* **267**, 133–164. 14

[64] Bardou, M, Montembault, S, Giraud, V, et al. (2002) Excessive alcohol consumption favours high risk polyp or colorectal cancer occurrence among patients with adenomas: a case control study. *Gut* **50**, 38–42. 14

[65] Vioque, J, Barber, X, Bolumar, F, et al. (2008) Esophageal cancer risk by type of alcohol drinking and smoking: a case-control study in spain. *BMC Cancer* **8**, 221. 14

[66] Pleasance, E. D, Stephens, P. J, O'Meara, S, et al. (2010) A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190. 14

[67] Barone-Adesi, F, Chapman, R. S, Silverman, D. T, et al. (2012) Risk of lung cancer associated with domestic use of coal in Xuanwei, China: retrospective cohort study. *BMJ* **345**, e5414. 14

[68] Fearon, E. R. (1997) Human Cancer Syndromes: Clues to the Origin and Nature of Cancer. *Science* **278**, 1043–1050. 14

[69] Lichtenstein, P, Holm, N. V, Verkasalo, P. K, et al. (2000) Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland. *New England Journal of Medicine* **343**, 78–85. 14

[70] Tomasetti, C & Vogelstein, B. (2015) Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, 78–81. 14

[71] Hanahan, D & Weinberg, R. A. (2011) Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674. 14, 15, 16

[72] Knudson, A. G. (1971) Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America* **68**, 820–823. 14

[73] Fearon, E. R & Vogelstein, B. (1990) A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767. 14

[74] Louhimo, R & Hautaniemi, S. (2011) CNAmet: an R package for integrating copy number, methylation and expression data. *Bioinformatics* p. btr019. 14

[75] Huang, Q, Whitington, T, Gao, P, et al. (2014) A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding. *Nature Genetics* **46**, 126–135. 14

[76] Melton, C, Reuter, J. A, Spacek, D. V, et al. (2015) Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nature Genetics* **47**, 710–716. 14

[77] Szczyrba, J, Löprich, E, Wach, S, et al. (2010) The MicroRNA Profile of Prostate Carcinoma Obtained by Deep Sequencing. *Molecular Cancer Research* **8**, 529–538. 14

[78] Jansson, M. D & Lund, A. H. (2012) MicroRNA and cancer. *Molecular Oncology* **6**, 590–610. 14

[79] Reva, B, Antipin, Y, & Sander, C. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Research* **39**, e118. 15

[80] Yamada, Y, Banno, Y, Yoshida, H, et al. (2006) Catalytic inactivation of human phospholipase D2 by a naturally occurring Gly901asp mutation. *Archives of Medical Research* **37**, 696–699. 15

[81] Baylin, S. B. (2005) DNA methylation and gene silencing in cancer. *Nature Clinical Practice Oncology* **2**, S4–S11. 15

[82] Ehrlich, M. (2009) DNA hypomethylation in cancer cells. *Epigenomics* **1**, 239–259. 15

[83] Esteller, M. (2007) Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature Reviews Genetics* **8**, 286–298. 15

[84] Pruitt, K, Zinn, R. L, Ohm, J. E, et al. (2006) Inhibition of SIRT1 reactivates silenced cancer genes without loss of promoter DNA hypermethylation. *PLoS Genetics* **2**, e40. 15

[85] Fraga, M. F, Ballestar, E, Villar-Garea, A, et al. (2005) Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. *Nature Genetics* **37**, 391–400. 15

[86] Kreeger, P. K & Lauffenburger, D. A. (2010) Cancer systems biology: a network modeling perspective. *Carcinogenesis* **31**, 2–8. 15, 39

[87] Jones, R, Ruas, M, Gregory, F, et al. (2007) A CDKN2a Mutation in Familial Melanoma that Abrogates Binding of p16ink4a to CDK4 but not CDK6. *Cancer Research* **67**, 9134–9141. 15

[88] Chen, J, Wang, Y, Shen, B, et al. (2013) Molecular Signature of Cancer at Gene Level or Pathway Level? Case Studies of Colorectal Cancer and Prostate Cancer Microarray Data. *Computational and Mathematical Methods in Medicine* **2013, 2013**, e909525. 15

[89] van 't Veer, L. J, Dai, H, van de Vijver, M. J, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536. 15, 37

[90] Zhao, F, Siu, M. K. Y, Jiang, L, et al. (2014) Overexpression of forkhead box protein M1 (FOXM1) in ovarian cancer correlates with poor patient survival and contributes to paclitaxel resistance. *PloS One* **9**, e113478. 15

[91] Pasanen, A. K, Haapasaari, K.-M, Peltonen, J, et al. (2013) Cell cycle regulation score predicts relapse-free survival in non-germinal centre diffuse large B-cell lymphoma patients treated by means of immunochemotherapy. *European Journal of Haematology* **91**, 29–36. 15

[92] Sánchez-Beato, M, Sánchez-Aguilera, A, & Piris, M. A. (2003) Cell cycle deregulation in B-cell lymphomas. *Blood* **101**, 1220–1235. 15

[93] Collins, K, Jacks, T, & Pavletich, N. P. (1997) The cell cycle and cancer. *Proceedings of the National Academy of Sciences* **94**, 2776–2778. 15

[94] Ortega, S, Malumbres, M, & Barbacid, M. (2002) Cyclin D-dependent kinases, INK4 inhibitors and cancer. *Biochimica et Biophysica Acta* **1602**, 73–87. 15

[95] Sherr, C. J & McCormick, F. (2002) The RB and p53 pathways in cancer. *Cancer Cell* **2**, 103–112. 15

[96] Hahn, W. C & Weinberg, R. A. (2002) Modelling the molecular circuitry of cancer. *Nature Reviews Cancer* **2**, 331–341. 15

[97] Dhillon, A. S, Hagan, S, Rath, O, et al. (2007) MAP kinase signalling pathways in cancer. *Oncogene* **26**, 3279–3290. 16

[98] Ideker, T, Galitski, T, & Hood, L. (2001) A NEW APPROACH TO DECODING LIFE: Systems Biology. *Annual Review of Genomics and Human Genetics* **2**, 343–372. 16

[99] Regenmortel, M. H. V. V. (2004) Reductionism and complexity in molecular biology. *EMBO reports* **5**, 1016–1020. 16

[100] Kreeger, P. K & Lauffenburger, D. A. (2010) Cancer systems biology: a network modeling perspective. *Carcinogenesis* **31**, 2–8. 16

[101] Ahn, A. C, Tewari, M, Poon, C.-S, et al. (2006) The Limits of Reductionism in Medicine: Could Systems Biology Offer an Alternative? *PLoS Medicine* **3**, e208. 16

[102] Edgar, R, Domrachev, M, & Lash, A. E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30**, 207–210. 16, 21

[103] Mehta, S, Shelling, A, Muthukaruppan, A, et al. (2010) Predictive and prognostic molecular markers for cancer medicine. *Therapeutic Advances in Medical Oncology* **2**, 125–148. 16

[104] Curtis, C, Shah, S. P, Chin, S.-F, et al. (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352. 17

[105] Stratton, M. R, Campbell, P. J, & Futreal, P. A. (2009) The cancer genome. *Nature* **458**, 719–724. 17

[106] Akavia, U. D, Litvin, O, Kim, J, et al. (2010) An Integrated Approach to Uncover Drivers of Cancer. *Cell* **143**, 1005–1017. 17

[107] Livshits, A, Git, A, Fuks, G, et al. (2015) Pathway-based personalized analysis of breast cancer expression data. *Molecular Oncology* **9**, 1471–1483. 17

[108] Drier, Y, Sheffer, M, & Domany, E. (2013) Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences* **110**, 6388–6393. 17, 24, 27, 39

[109] Aylin, P, Bottle, A, & Majeed, A. (2007) Use of administrative data or clinical databases as predictors of risk of death in hospital: comparison of models. *BMJ* **334**, 1044. 17

[110] Chen, P, Huhtinen, K, Kaipio, K, et al. (2015) Identification of Prognostic Groups in High-Grade Serous Ovarian Cancer Treated with Platinum-Taxane Chemotherapy. *Cancer Research* **75**, 2987–2998. 17, 18, 19

[111] Louhimo, R. (2015) Ph.D. thesis (University of Helsinki). 17, 24

[112] Siegel, R. L, Miller, K. D, & Jemal, A. (2015) Cancer statistics, 2015. *CA: a cancer journal for clinicians* **65**, 5–29. 18

[113] Sørlie, T, Perou, C. M, Tibshirani, R, et al. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences* **98**, 10869–10874. 18

[114] Gazinska, P, Grigoriadis, A, Brown, J. P, et al. (2013) Comparison of basal-like triple-negative breast cancer defined by morphology, immunohistochemistry and transcriptional profiles. *Modern Pathology* **26**, 955–966. 18

[115] Oakman, C, Viale, G, & Di Leo, A. (2010) Management of triple negative breast cancer. *Breast* **19**, 312–321. 18

[116] Kreike, B, van Kouwenhove, M, Horlings, H, et al. (2007) Gene expression profiling and histopathological characterization of triple-negative/basal-like breast carcinomas. *Breast Cancer Research* **9**, R65. 18

[117] Lumachi, F, Brunello, A, Maruzzo, M, et al. (2013) Treatment of estrogen receptor-positive breast cancer. *Current Medicinal Chemistry* **20**, 596–604. 18

[118] Jezequel, P, Loussouarn, D, Guerin-Charbonnel, C, et al. (2015) Gene-expression molecular subtyping of triple-negative breast cancer tumours: importance of immune response. *Breast Cancer Research* **17**, 43. 18

[119] Armstrong, D. K. (2002) Relapsed Ovarian Cancer: Challenges and Management Strategies for a Chronic Disease. *Oncologist* **7**, 20–28. 18

[120] Schwarz, R. F, Ng, C. K. Y, Cooke, S. L, et al. (2015) Spatial and Temporal Heterogeneity in High-Grade Serous Ovarian Cancer: A Phylogenetic Analysis. *PLoS Medicine* **12**, e1001789. 18

[121] Pal, T, Permuth-Wey, J, Betts, J. A, et al. (2005) BRCA1 and BRCA2 mutations account for a large proportion of ovarian carcinoma cases. *Cancer* **104**, 2807–2816. 18

[122] Pfreundschuh, M, Trümper, L, Osterborg, A, et al. (2006) CHOP-like chemotherapy plus rituximab versus CHOP-like chemotherapy alone in young patients with good-prognosis diffuse large-B-cell lymphoma: a randomised controlled trial by the MabThera International Trial (MInT) Group. *The Lancet. Oncology* **7**, 379–391. 19

[123] Coiffier, B, Lepage, E, Briere, J, et al. (2002) CHOP chemotherapy plus rituximab compared with CHOP alone in elderly patients with diffuse large-B-cell lymphoma. *New England Journal of Medicine* **346**, 235–242. 19

[124] Klyuchnikov, E, Bacher, U, Kroll, T, et al. (2014) Allogeneic hematopoietic cell transplantation for diffuse large B cell lymphoma: who, when and how? *Bone Marrow Transplantation* **49**, 1–7. 19

[125] Vaidya, R & Witzig, T. E. (2014) Prognostic factors for diffuse large b-cell lymphoma in the r(x)chop era. *Annals of Oncology* **25**, 2124. 19

[126] Lenz, G & Staudt, L. M. (2010) Aggressive lymphomas. *New England Journal of Medicine* **362**, 1417–1429. 19

[127] Su, Y, Nielsen, D, Zhu, L, et al. (2013) Gene selection and cancer type classification of diffuse large-b-cell lymphoma using a bivariate mixture model for two-species data. *Human Genomics* **7**, 2. 19

[128] Wilson, W. H. (2013) Treatment strategies for aggressive lymphomas: what works? *ASH Education Program Book* **2013**, 584–590. 19

[129] Alizadeh, A, Elsen, M, Davis, R, et al. (2000) Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511. 19

[130] Lenz, G & Staudt, L. M. (2010) Aggressive Lymphomas. *New England Journal of Medicine* **362**, 1417–1429. 19

[131] Roschewski, M, Staudt, L. M, & Wilson, W. H. (2014) Diffuse large B-cell lymphoma-treatment approaches in the molecular era. *Nature Reviews Clinical Oncology* **11**, 12–23. 19, 36

[132] Rossi, D, Ciardullo, C, & Gaidano, G. (2013) Genetic aberrations of signaling pathways in lymphomagenesis: revelations from next generation sequencing studies. *Seminars in Cancer Biology* **23**, 422–430. 19, 36

[133] Morin, R. D, Mungall, K, Pleasance, E, et al. (2013) Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing. *Blood* **122**, 1256–1265. 21

[134] Irizarry, R. A, Hobbs, B, Collin, F, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264. 21

[135] Ovaska, K, Laakso, M, Haapa-Paananen, S, et al. (2010) Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Medicine* **2**, 65. 22, 23

[136] Icay, K, Chen, P, Cervera, A, et al. (2016) Sepia: Rna and small rna sequence processing, integration, and analysis. *BioData Mining* **9**, 20. 22

[137] Khatri, P, Sirota, M, & Butte, A. J. (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Computational Biology* **8**. 22, 23, 41

[138] Huang, D. W, Sherman, B. T, & Lempicki, R. A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**, 44–57. 22

[139] Tarca, A. L, Draghici, S, Khatri, P, et al. (2008) A Novel Signaling Pathway Impact Analysis (SPIA). *Bioinformatics*. 23, 35

[140] Laakso, M & Hautaniemi, S. (2010) Integrative platform to translate gene sets to networks. *Bioinformatics* **26**, 1802–1803. 23

[141] Cerami, E. G, Gross, B. E, Demir, E, et al. (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research* **39**, D685–D690. 23

[142] Ahn, T, Lee, E, Huh, N, & Park, T. (2014) Personalized identification of altered pathways in cancer using accumulated normal tissue data. *Bioinformatics* **30**, i422–i429. 24, 27, 39

[143] Goel, M. K, Khanna, P, & Kishore, J. (2010) Understanding survival analysis: Kaplan-Meier estimate. *International Journal of Ayurveda Research* **1**, 274–278. 25

[144] Altman, D. G. (1990) *Practical Statistics for Medical Research*. (CRC Press). 25, 26

[145] Li, D, Fu, B, Wang, Y, et al. (2015) Percolation transition in dynamical traffic network with evolving critical bottlenecks. *Proceedings of the National Academy of Sciences* **112**, 669–672. 27, 28

[146] Ertych, N, Stolz, A, Stenzinger, A, et al. (2014) Increased microtubule assembly rates influence chromosomal instability in colorectal cancer cells. *Nature Cell Biology* **16**, 779–791. 31

[147] Justilien, V, Jameison, L, Der, C. J, et al. (2011) Oncogenic activity of Ect2 is regulated through protein kinase C iota-mediated phosphorylation. *Journal of Biological Chemistry* **286**, 8149–8157. 31

[148] Evers, L, Perez-Mancera, P. A, Lenkiewicz, E, et al. (2014) STAG2 is a clinically relevant tumor suppressor in pancreatic ductal adenocarcinoma. *Genome Medicine* **6**, 9. 31

[149] Ito, Y, Miyoshi, E, Sasaki, N, et al. (2004) Polo-like kinase 1 overexpression is an early event in the progression of papillary carcinoma. *British Journal of Cancer* **90**, 414–418. 31

[150] McInnes, C & Wyatt, M. D. (2011) PLK1 as an oncology target: current status and future potential. *Drug Discovery Today* **16**, 619–625. 31

[151] Sebastian, M, Reck, M, Waller, C. F, et al. (2010) The efficacy and safety of BI 2536, a novel Plk-1 inhibitor, in patients with stage IIIB/IV non-small cell lung cancer who had relapsed after, or failed, chemotherapy: results from an open-label, randomized phase II clinical trial. *Journal of Thoracic Oncology* **5**, 1060–1067. 31

[152] Hofheinz, R.-D, Al-Batran, S.-E, Hochhaus, A, et al. (2010) An open-label, phase I study of the polo-like kinase-1 inhibitor, BI 2536, in patients with advanced solid tumors. *Clinical Cancer Research* **16**, 4666–4674. 31

[153] Mehta, R. J, Jain, R. K, Leung, S, et al. (2012) FOXA1 is an independent prognostic marker for ER-positive breast cancer. *Breast Cancer Research and Treatment* **131**, 881–890. 33

[154] Sieuwerts, A. M, Look, M. P, Meijer-van Gelder, M. E, et al. (2006) Which cyclin E prevails as prognostic marker for breast cancer? Results from a retrospective study involving 635 lymph node-negative breast cancer patients. *Clinical Cancer Research* **12**, 3319–3328. 33

[155] Signoretti, S, Di Marcotullio, L, Richardson, A, et al. (2002) Oncogenic role of the ubiquitin ligase subunit Skp2 in human breast cancer. *The Journal of Clinical Investigation* **110**, 633–641. 33

[156] Chan, C.-H, Morrow, J. K, Li, C.-F, et al. (2013) Pharmacological inactivation of Skp2 SCF ubiquitin ligase restricts cancer stem cell traits and cancer progression. *Cell* **154**, 556–568. 33

[157] Wang, B, Mezlini, A. M, Demir, F, et al. (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods* **11**, 333–337. 34

[158] Formosa, A, Markert, E. K, Lena, A. M, et al. (2014) MicroRNAs, miR-154, miR-299-5p, miR-376a, miR-376c, miR-377, miR-381, miR-487b, miR-485-3p, miR-495 and miR-654-3p, mapped to the 14q32.31 locus, regulate proliferation, apoptosis, migration and invasion in metastatic prostate cancer cells. *Oncogene* **33**, 5173–5182. 34

[159] Zheng, B, Liang, L, Huang, S, et al. (2012) MicroRNA-409 suppresses tumour cell invasion and metastasis by directly targeting radixin in gastric cancers. *Oncogene* **31**, 4509–4516. 34

[160] Weng, C, Dong, H, Chen, G, et al. (2012) miR-409-3p inhibits HT1080 cell proliferation, vascularization and metastasis by targeting angiogenin. *Cancer Letters* **323**, 171–179. 34

[161] Okamoto, K, Ishiguro, T, Midorikawa, Y, et al. (2012) miR-493 induction during carcinogenesis blocks metastatic settlement of colon cancer cells in liver. *EMBO journal* **31**, 1752–1763. 34

[162] Sakai, H, Sato, A, Aihara, Y, et al. (2014) MKK7 mediates miR-493-dependent suppression of liver metastasis of colon cancer cells. *Cancer Science* **105**, 425–430. 34

[163] Guo, H, Ingolia, N. T, Weissman, J. S, et al. (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**, 835–840. 35

[164] Thomas, S. J, Snowden, J. A, Zeidler, M. P, et al. (2015) The role of JAK/STAT signalling in the pathogenesis, prognosis and treatment of solid tumours. *British Journal of Cancer* **113**, 365–371. 35

[165] Diaz, T, Navarro, A, Ferrer, G, et al. (2011) Lestaurtinib Inhibition of the JAK/STAT Signaling Pathway in Hodgkin Lymphoma Inhibits Proliferation and Induces Apoptosis. *PLoS ONE* **6**, e18856. 35

[166] Furqan, M, Mukhi, N, Lee, B, et al. (2013) Dysregulation of JAK-STAT pathway in hematological malignancies and JAK inhibitors for clinical application. *Biomarker Research* **1**, 5.

[167] Niemann, C. U & Wiestner, A. (2013) B-cell receptor signaling as a driver of lymphoma development and evolution. *Seminars in Cancer Biology* **23**, 410–421.

[168] Gramling, M. W & Eischen, C. M. (2012) Suppression of Ras/Mapk pathway signaling inhibits Myc-induced lymphomagenesis. *Cell Death and Differentiation* **19**, 1220–1227.

[169] Elenitoba-Johnson, K. S. J, Jenson, S. D, Abbott, R. T, et al. (2003) Involvement of multiple signaling pathways in follicular lymphoma transformation: p38-mitogen-activated protein kinase as a target for therapy. *Proceedings of the National Academy of Sciences* **100**, 7259–7264.

[170] Han, J & Shen, Q. (2012) Targeting γ-secretase in breast cancer. *Breast Cancer : Targets and Therapy* **4**, 83–90.

[171] Haapasalo, A & Kovacs, D. M. (2011) The Many Substrates of Presenilin/γ-Secretase. *Journal of Alzheimer's Disease* **25**, 3–28.

[172] Vicario, A, Kisiswa, L, Tann, J. Y, et al. (2015) Neuron-type-specific signaling by the p75ntr death receptor is regulated by differential proteolytic cleavage. *J Cell Sci* **128**, 1507–1517.

[173] DeFreitas, M. F, McQuillen, P. S, & Shatz, C. J. (2001) A novel p75ntr signaling pathway promotes survival, not death, of immunopurified neocortical subplate neurons. *The Journal of Neuroscience* **21**, 5121–5129.

[174] Keino-Masu, K, Masu, M, Hinck, L, et al. (1996) Deleted in Colorectal Cancer (DCC) Encodes a Netrin Receptor. *Cell* **87**, 175–185.

[175] Forcet, C, Stein, E, Pays, L, et al. (2002) Netrin-1-mediated axon outgrowth requires deleted in colorectal cancer-dependent MAPK activation. *Nature* **417**, 443–447.

[176] Liu, X, Wang, Y, Ji, H, et al. (2016) Personalized characterization of diseases using sample-specific networks. *Nucleic Acids Research* **44**, e164–e164.

[177] Pawson, T & Linding, R. (2008) Network medicine. *FEBS letters* **582**, 1266–1270.

[178] Hofree, M, Shen, J. P, Carter, H, et al. (2013) Network-based stratification of tumor mutations. *Nature Methods* **10**, 1108–1115.

[179] Valencia, A & Hidalgo, M. (2012) Getting personalized cancer genome analysis into the clinic: the challenges in bioinformatics. *Genome Medicine* **4**, 61.

[180] Greaves, M & Maley, C. C. (2012) Clonal evolution in cancer. *Nature* **481**, 306–313. 01228.

[181] Shackleton, M, Quintana, E, Fearon, E. R, & Morrison, S. J. (2009) Heterogeneity in Cancer: Cancer Stem Cells versus Clonal Evolution. *Cell* **138**, 822–829. 00799.

[182] Haynes, W. A, Higdon, R, Stanberry, L, et al. (2013) Differential Expression Analysis for Pathways. *PLoS Computational Biology* **9**, e1002967.

[183] Nam, S, Chang, H. R, Kim, K.-T, et al. (2014) PATHOME: an algorithm for accurately detecting differentially expressed subpathways. *Oncogene* **33**, 4941–4951. 40

[184] Calin, G. A & Croce, C. M. (2006) MicroRNA signatures in human cancers. *Nature Reviews Cancer* **6**, 857–866. 41

[185] Iorio, M. V & Croce, C. M. (2012) MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review. *EMBO molecular medicine* **4**, 143–159. 41

[186] Iqbal, J, Shen, Y, Huang, X, et al. (2015) Global microRNA expression profiling uncovers molecular markers for classification and prognosis in aggressive B-cell lymphoma. *Blood* **125**, 1137–1145. 41

[187] Alencar, A. J, Malumbres, R, Kozloski, G. A, et al. (2011) MicroRNAs are independent predictors of outcome in diffuse large B-cell lymphoma patients treated with R-CHOP. *Clinical Cancer Research* **17**, 4125–4135. 41

[188] Gingeras, T. R. (2007) Origin of phenotypes: Genes and transcripts. *Genome Research* **17**, 682–690. 41