Multi-omics integrative analyses for decision support systems in personalized cancer treatment

Dissertation

zur Erlangung des Grades des Doktors der Naturwissenschaften (Dr. rer. nat.)
an der Fakultät für Mathematik und Informatik der Universität des Saarlandes

von
Lara Kristina Schneider

Saarbrücken, 2020

Tag des Kolloquiums:	**10. Juni 2020**
Dekan der Fakultät:	**Prof. Dr. Thomas Schuster**

Prüfungsausschuss:

Vorsitzende:	**Prof. Dr. Verena Wolf**
Erstgutachter:	**Prof. Dr. Hans-Peter Lenhof**
Zweitgutachter:	**Prof. Dr. Andreas Keller**
Wissenschaftliche Mitarbeiterin:	**Dr. Christina Backes**

Saarland University

Center for Bioinformatics

Faculty for Mathematics and Informatics

Saarbrücken Graduate School of Computer Science

Dissertation

Multli-omics integrative analyses for decision support systems in personalized cancer treatment

by

Lara Kristina Schneider

Saarbrücken, 2020

Thesis advisor

Prof. Dr. Hans-Peter Lenhof

Second examiner

Prof. Dr. Andreas Keller

Abstract

Cancer is a heterogeneous class of complex diseases that are characterized by unlimited proliferation of malignant cells. Cancer is caused by a complex interplay of (epi-)genetic aberrations that strongly differ between tumors. This heterogeneity makes cancer difficult to treat. Personalized cancer treatment aims at selecting the most suitable (combination of) drugs to treat a specific tumor based on its genetic and molecular characteristics. Data from modern high-throughput experimental technologies have greatly improved the resolution at which tumor-driving factors can be determined across levels of cellular regulation. In order to extract disease- and treatment-relevant information from this mass of complex, high-dimensional, and noisy data sets, robust statistical and computational tools and methods are required.

In this thesis, we present a comprehensive tool suite for cancer treatment decision support and translational research. The encompassed tools provide rich functionality for the genetic and molecular characterization of tumors with an emphasis on deregulated biological processes and the identification of disease-driving regulatory key players. The integrative analysis of multi-omics data sets with *a priori* knowledge from clinical practice guidelines and relevant medical, pharmacological, and biological databases covers a variety of research scenarios from biomarker identification to the personalized assessment of various types of treatment options including standard-of-care targeted drugs, candidates for drug repositioning and immunotherapy.

First, we present several tools for the statistical analysis of multi-omics data that can be broadly applied, e.g. for molecular characterization of a sample of interest or for biomarker identification. These tools include GeneTrail2 and RegulatorTrail. GeneTrail2 is a web service for the statistical analysis of molecular signatures that provides various types of enrichment analyses for the identification of deregulated pathways. While GeneTrail2 focuses on the assessment of aberrant biological pathways and signaling cascades, RegulatorTrail aims at the identification of those transcriptional regulators that seem to have a high impact on these pathogenic processes. To this end, RegulatorTrail provides numerous methods. As one of these methods, we propose REGulator-Gene Association Enrichment (REGGAE), which is based on the combination of regulator-target interactions and enrichment analysis.

While the above-mentioned tools and methods are designed for general purposes, we specifically focus on personalized medicine and translational research with DrugTargetInspector (DTI). DTI provides rich functionality for the assessment of molecular drug targets, putative target pathways and corresponding drugs based on the integrative analysis of tumor-specific omics data sets.

Finally, we present ClinOmicsTrail^{bc}, an interactive visual analytics tool for breast cancer treatment stratification. ClinOmicsTrail^{bc} supports clinicians by providing a thorough assessment of standard-of-care targeted drugs, candidates for drug repositioning, and immunotherapeutic approaches, including checkpoint inhibitors and personalized cancer vaccines.

In summary, this work presents novel methods and computational tools for the integrative analysis of multi-omics data for translational research and clinical decision support, assisting researchers and clinicians in finding the best possible treatment options in a deeply personalized way.

Summary

Cancer is a heterogeneous class of diseases caused by the complex interplay of (epi-)genetic aberrations and is difficult to treat due to its heterogeneity. In this thesis, we present a tool suite of novel methods and computational tools for the genetic and molecular characterization of tumors to support decision making in personalized cancer treatments. The first tool included in this tool suite is GeneTrail2, a web service for the statistical analysis of molecular signatures. While GeneTrail2 focuses on the evaluation of aberrant biological pathways and signal cascades, RegulatorTrail identifies those transcriptional regulators that appear to have a high impact on these pathogenic processes. With DrugTargetInspector (DTI), we focus specifically on personalized medicine and translational research. DTI offers comprehensive functions for the evaluation of target molecules, the associated signaling cascades, and the corresponding drugs. Finally, we present ClinOmicsTrail^{bc}, an analysis tool for stratification of breast cancer treatments. ClinOmicsTrail^{bc} supports clinicians with a comprehensive evaluation of on- and off-label drugs as well as immune therapies.

Zusammenfassung

Krebs ist eine heterogene Klasse von Erkrankungen, die durch ein komplexes Zusammenspiel von (epi-)genetischen Aberrationen verursacht wird und sich aufgrund ihrer Heterogenität nur schwer behandeln lässt. In dieser Arbeit präsentieren wir eine Reihe neuer Methoden und Berechnungswerkzeuge für die genetische und molekulare Charakterisierung von Tumoren zur Entscheidungsunterstützung bei personalisierten Krebsbehandlungen.

Diese Berechnungswerkzeuge beinhalten unter anderem GeneTrail2, einen Webservice für die statistische Analyse molekularer Signaturen. Während GeneTrail2 sich auf die Bewertung von aberranten biologischen Pfaden und Signalkaskaden konzentriert, identifiziert RegulatorTrail jene Transkriptionsregulatoren, die einen hohen Einfluss auf diese pathogenen Prozesse zu haben scheinen. Mit DrugTargetInspector (DTI) fokussieren wir uns speziell auf die personalisierte Medizin und die translationale Forschung. DTI bietet umfangreiche Funktionen für die Bewertung von Zielmolekülen, den dazugehörigen Signalkaskaden und entsprechenden Medikamenten. Schließlich präsentieren wir ClinOmicsTrail^{bc}, ein Analysetool zur Stratifizierung von Brustkrebsbehandlungen. ClinOmicsTrail^{bc} unterstützt Kliniker durch eine umfassende Bewertung von On- und Off-Label-Medikamenten und Immuntherapien.

Acknowledgments

There are many people from whose support this thesis has benefited greatly and who deserve special acknowledgment here. First and foremost, I would like to thank my supervisor Prof. Dr. Hans-Peter Lenhof for his continuous support and scientific input.

As computational biology is a truly interdisciplinary field of research, this work would not have been possible without the guidance and contributions by our collaboration partners. Many thanks to the following professors and their research groups: Prof. Dr. Andreas Keller, Prof. Dr. Eckart Meese, Prof. Dr. Norbert Graf, Prof. Dr. Oliver Kohlbacher, Prof. Dr. Markus Wallwiener, Prof. Dr. Andreas Hartkopf, Prof. Dr. Andreas Hildebrandt, Prof. Dr. Michael Kaufmann, Prof. Dr. Stefan Tenzer, Prof. Dr. Hanno Huwer, and Prof. Dr. Ulrich Keller.

Moreover, I am grateful to all my current and former colleagues at CBI and MPII, as well as the students I had the pleasure of working with over the years. Thank you for creating such a friendly working environment (in alphabetical order): Anke King, Anna Hake, Anne Hildebrandt, Carolin Mayer, Christina Backes, Daniel Stöckel, David Porubsky, Dilip Durai, Fabian Kern, Fabian Müller, Fatemeh Behjati, Felipe Albrecht, Florian Schmidt, Georges Schmartz, Jana Ebler, Jérémy Amand, Julia Jauß, Kerstin Lenhof, Kristina Thedinga, Lea Eckhart, Lisa Handl, Marc Hellmuth, Maryam Ghareghani, Matthias Dietzen, Matthias Döring, Max Jakob, Michael Scherer, Mikko Rautiainen, Mustafa Kahraman, Nadja Liddy Grammes, Nico Gerstner, Nico Pfeifer, Nora Speicher, Olga Kalinina, Patrick Trampert, Paula Linh Kramer, Peter Ebert, Prabhav Kalaghatgi, Rebecca Serra Mari, Sivarajan Karunanithi, Sven Hafeneger, Tim Kehl, Tobias Fehlman, Tobias Marschall, Tomas Bastys, Thorsten Will, and Valentina Galata.

I am particularly thankful to Pia Scherer-Geiß, Karin Jostock, Francoise Laroppe, and Nadine Willhelm for their administrative and moral support throughout the years.

Last, but definitely not least, I would like to express my deepest gratitude to Karsten Knuth and to my parents, who supported and encouraged me throughout the years.

Contents

1

Introduction

Cancer is a class of complex diseases characterized by uncontrolled proliferation of cells that tend to invade surrounding tissue and metastasize to other sites of the body [1]. By the year 2030, two out of three elderly people are expected to be diagnosed with cancer [2]. As incidence rates of cancer are strongly correlated with age [3], a prolonged life expectancy and an aging population will result in a continuously increasing number of cancer patients in the future. Although the progress in science and medicine has improved the treatment of cancer over the last decades [4], cancer is still the second leading cause of death in Germany and other Western industrialized countries (cf. **Figure 1.1**).

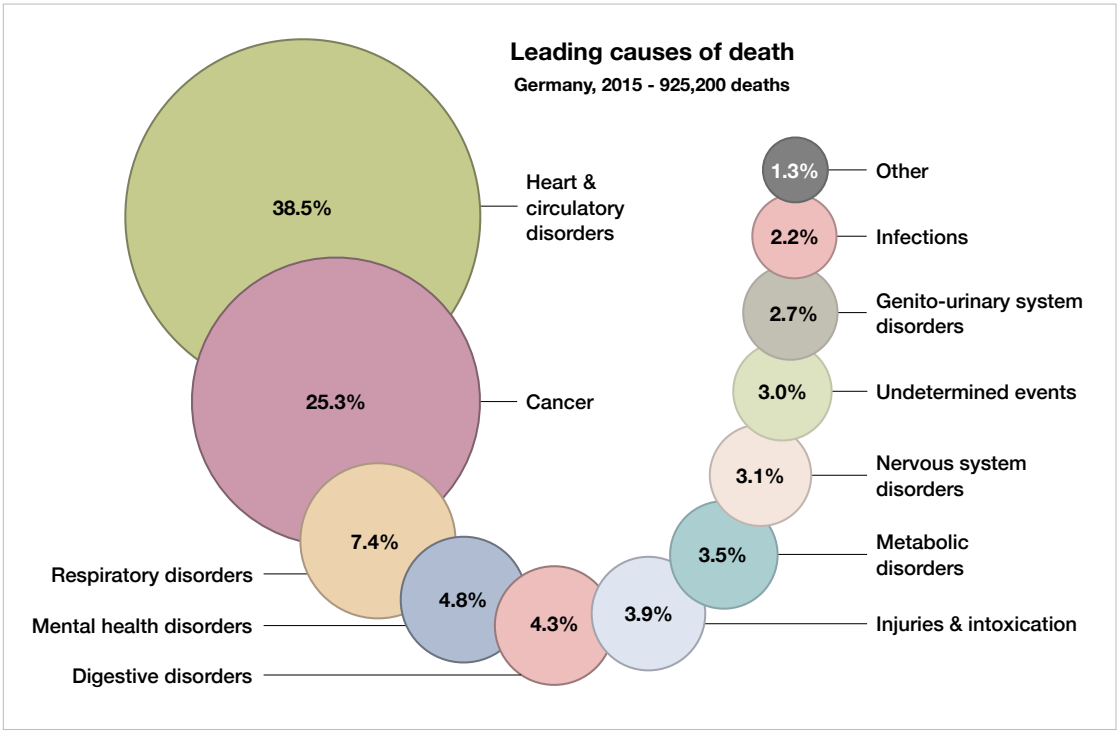


Figure 1.1 **Leading causes of death in Germany 2015**. Based on a total of 925,200 cases, the fraction of the eleven most common causes of death are displayed. Data obtained from the German Federal Statistical Office [5].

The reasons why cancer treatment is still a grand challenge can be found in the origins and traits of this complex class of diseases. The transformation from normal human cells into cancer cells occurs via the acquisition of several capabilities enabling malignant growth. High mutation rates combined with uncontrolled proliferation lead to a Darwinian-like cellular evolution that fosters tumor heterogeneity. This diversity is also reflected in strongly varying treatment responses in

tumors of the same type or even subtype and hence has to be accounted for in a personalized treatment decision-making process [6].

The idea of personalized medicine *per se* is not new. The field of medicine has always been personalized in such a way that physicians have strived to determine the underlying causes of their patients' diseases and to treat them accordingly. It is conveyed that already around 400 BC, the Greek physician Hippocrates assessed the composition of four distinct types of body fluids black and yellow bile, phlegm, and blood (the 'four humors') before treating a patient [7].

Over the last decades, there were numerous endeavors to identify treatment-relevant biomarkers that also led to the concept of companion diagnostics. Companion diagnostics are specific molecular, genetic, or imaging tests that assess the status of key tumor-driving genes or proteins, which inform the applicability of specific drugs. For example, the treatment of colon cancer with cetuximab requires the determination of the mutation status of the Kirsten rat sarcoma viral oncogene homolog (KRAS). KRAS is a proto-oncogene that oftentimes carries an activating mutation rendering KRAS independent from upstream activation by the epidermal growth factor receptor (EGFR), the molecular drug target of cetuximab. Hence, treatment with cetuximab is only eligible for patients without the mutation [8]. Nowadays, an increasing number of companion diagnostics for cancer treatment stratification are approved by the U.S. Food and Drug Administration (FDA). However, in many cases, the sole consideration of single or a few biomarkers might not suffice to comprehensively identify the specific tumor's driving factors and hence predict treatment outcome. Oncogenic mutations in the v-raf murine sarcoma viral oncogene homolog B (BRAF) are illustrative for this, as they are predictive for BRAF inhibitor response (e.g., for treatment with vemurafenib or dabrafenib) in melanoma [9, 10], but not necessarily in other cancer types. [11].

Although there is anecdotal evidence of super responders that experienced long-term remission after personalized treatment [12], its broad clinical benefit remains to be shown in clinical trials [13]. The complex biology of altered signaling cascades driving a tumor is likely to explain why there are currently only a few classes of patients benefiting from precision oncology. Hence, in order to comprehensively characterize and stratify a given tumor, besides of the sole consideration of individual actionable mutations, the complex molecular 'circuitry' of altered genes and proteins and their interdependencies also need to be taken into account to obtain a holistic view on disease-driving mechanisms.

With an evolving understanding of cancer biology and the increased accessibility of high-throughput technologies, the breadth and depth at which tumor-driving factors can be determined on a genetic and molecular level in clinically relevant time frames have greatly improved. Biotechnological high-throughput methods of various types enable fine-grained measurements of a tumor's (epi-)genome and transcriptome, microRNAs, proteins, and metabolites ('omics data') in the cell.

The integrative analysis of all of these multi-faceted data sets has the potential to significantly advance personalized medicine [14]. However, the sheer amount of high-dimensional and noisy data and the complexity of molecular interactions and dependencies make the manual identification of all treatment-relevant pieces of information a futile task. Hence, easy-to-use, yet powerful bioinformatics tools and clinical decision support systems are required that integrate clinical and multi-omics data sets of a patient under investigation with *a priori* knowledge from various biological, medical, and pharmacological databases and reference data sets. Such

analyses have to be conducted in a transparent and reproducible manner. The obtained results should be provided in clear and easy-to-interpret visualizations that support clinicians in selecting the best suitable treatment options for their patients (cf. **Figure 1.2**).
Personalized cancer treatment on the basis of such a holistic assessment of key tumor driving genes and pathways has the potential to improve treatment responses and quality of life for patients. Moreover, the elucidation of specific tumor characteristics and how they inform drug sensitivity can promote drug development by revealing additional or alternative uses for drugs and drug candidates.

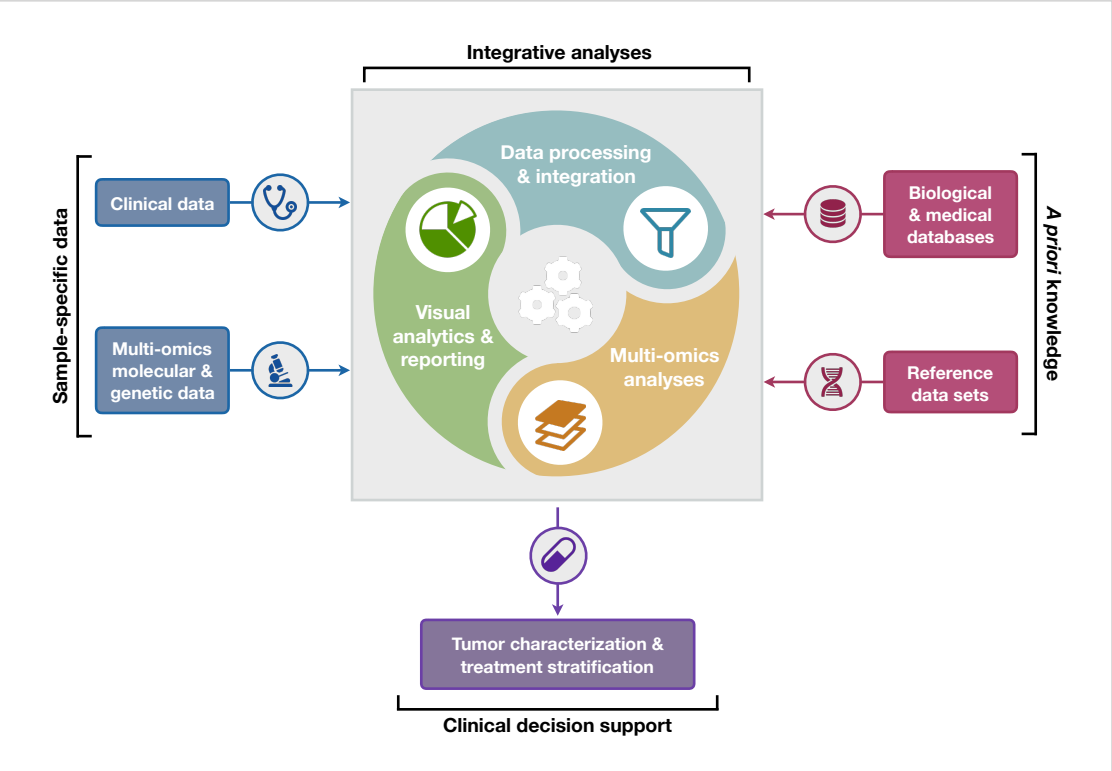


Figure 1.2 **Workflow of decision support for personalized cancer treatment using multi-omics integrative analyses.** The integrative analysis of a tumor sample's clinical and molecular data in combination with *a priori* knowledge from various biological and medical databases can be employed to characterize the given tumor, hence providing clinical decision support for treatment stratification. The icons in this figure were obtained from [15].

Thesis scope and outline

In this thesis, we present a comprehensive tool suite for cancer treatment decision support and translational research. The encompassed tools provide rich functionality for the genetic and molecular characterization of tumors with an emphasis on deregulated biological processes and the identification of disease-driving regulatory key players. The identified tumor characteristics are combined with *a priori* knowledge from clinical practice guidelines and relevant medical, pharmacological, and biological databases for a personalized assessment of various types of treatment options, including standard-of-care targeted drugs, candidates for drug repositioning, and immunotherapy.

Tumors can potentially contain a plethora of (epi-)genomic and transcriptomic aberrations that manifest in dysregulated activity patterns of a variety of biological pathways, thereby leading to the specific disease phenotype. However, the alterations present in a tumor are not all contributing equally to the disease. Hence, for the identification of suitable treatment options, disease-driving alterations and their effects on cancer-relevant signaling cascades have to be determined. Such an assessment has the potential to drastically advance the evaluation of personalized (combination) therapies for cancer while improving the interpretability of analysis results.

Our first scientific contribution in this context is the development of GeneTrail2, a web service for the statistical analysis of molecular signatures that identifies deregulated pathways by means of enrichment analyses. GeneTrail2 offers numerous statistical tests and a broad collection of biological pathways and functional gene sets, making it one of the most comprehensive web service for enrichment analyses to date. While GeneTrail2 is a general-purpose tool for the assessment of altered biological pathways and signaling cascades, the identification of the underlying key regulatory elements that promote pathological processes is crucial to gain mechanistic insights into complex diseases like cancer. One essential class of regulatory elements implicated in cancer development and progression are transcriptional regulators. Being mostly located at the end of signaling cascades and hence acting as the effectors of intracellular signal transduction, the activation states of transcriptional regulators can also be considered as a proxy for the activities of their belonging pathways. We therefore developed RegulatorTrail, a web service for the assessment of transcriptional regulators with respect to their impact on pathogenic processes. RegulatorTrail provides eight different methods to identify and prioritize influential regulators on the basis of epigenomics and transcriptomics data. As one of these methods, we propose REGulator-Gene Association Enrichment (REGGAE), a novel approach to prioritize transcriptional regulators based on the combination of regulator-target interactions with Gene Set Enrichment Analysis. Using REGGAE, we were, for example, able to mechanistically elucidate the role of the transcription factor TCF3 as a potential master regulator in blastemal Wilms tumors, an aggressive form of childhood nephroblastoma.

The tools and methods mentioned above can be used for the thorough characterization of tumors, which is an essential prerequisite for personalized cancer treatment. The identified characteristics can be investigated with respect to their potential impact on drug sensitivity to inform treatment selection. In order to support oncologists in making informed treatment decisions and to foster translational research, we have developed the interactive assistance tool DrugTargetInspector (DTI). DTI provides rich functionality for the assessment of molecular drug targets, putative target pathways, and corresponding drugs based on the integrative analysis of tumor-specific omics data sets. Molecular drug targets of recommended drugs for more than 30 cancer types and a wide range of potential candidates for drug repositioning can be investigated with respect to their deregulation status and the existence of potentially resistance-causing mutations. Furthermore, DTI offers functionality to determine and visually assess a drug target's effect on downstream processes. While DrugTargetInspector is applicable across cancer types and provides a focused evaluation of molecular drug targets with potential applications in drug repositioning, the successful implementation of personalized medicine in the clinical practice requires the holistic assessment of a large variety of factors that might (de-)sensitize a drug in a cancer type-specific manner. Clinical decision support tools need

to comprehensively assess and integrate various types of (epi-)genomic and transcriptomic aberrations with respect to their implications for drug sensitivity while keeping the provided information as concise as possible. Moreover, in order to gain acceptance and build trust with clinicians and patients, transparency of the analyses and interpretability of the results are essential factors. To this end, we developed ClinOmicsTrail^{bc}, an interactive visual analytics tool for breast cancer treatment stratification. ClinOmicsTrail^{bc} supports clinicians by providing a thorough assessment of standard-of-care targeted drugs, candidates for drug repositioning, and immunotherapeutic approaches, including checkpoint inhibitors and personalized cancer vaccines. To this end, our tool analyzes clinical markers and (epi-)genomics and transcriptomics data sets to identify and evaluate the tumor's key driver mutations, the overall mutational load, activity patterns of cancer-relevant pathways, drug-specific predictive biomarkers, the status of molecular drug targets, and pharmacogenomic effects. The breadth and depth of analyzes and visualizations offered by ClinOmicsTrail^{bc} make it a promising tool with the potential to noticeably advance precision medicine and clinical decision support in the near future.

This thesis consists of seven chapters and its remainder is structured as follows: **Chapter 2** provides information on the required biological background. This includes an overview of essential components and characteristics of cancer development and progression, as well as a discussion of current treatment options and the concept of *personalized medicine* as an approach to cancer treatment. Across all methods and tools proposed in this thesis, we utilize a variety of databases providing *a priori* knowledge from various domains and large-scale multi-omics data sets that help to pinpoint disease-causing alterations. **Chapter 3** introduces the most relevant of these databases and data sets and gives an overview of current high-throughput experimental techniques that can be used to capture the aberrations present in a specific tumor. In **Chapter 4**, we present several tools and methods for the identification of various types of predictive biomarkers for treatment stratification. These include deregulated biological processes and pathway activities, as well as transcriptional regulators. For each of the presented tools and web services, we provide a description of the underlying methods and demonstrate their capabilities in several case studies on a diverse range of cancer types. **Chapter 5** describes the interactive assistance tool DrugTargetInspector. This includes a description of the tool's comprehensive functionality, as well as examples demonstrating its applicability to support personalized medicine and drug repositioning on different types of omics data. **Chapter 6** focuses on ClinOmicsTrail^{bc}, our visual analytics tool for breast cancer treatment stratification. We provide descriptions of the tool's rich set of features and highlight their relevance for treatment decision support. Moreover, we present several case studies demonstrating how ClinOmicsTrail^{bc} may guide the breast cancer treatment selection process. Finally, **Chapter 7** concludes this thesis and provides directions for future work.

Due to the interdisciplinary nature of bioinformatics research projects, this thesis contains many results that are based on the joint effort of various researchers across several disciplines. In order to ensure transparency of contributions, respective sections will contain information boxes on author contributions and references to corresponding publications.

2

Biological Background

This chapter introduces the biological background relevant for this thesis. To this end, we first discuss the basic principles of the molecular biology of eukaryotic cells, followed by a description of (epi-)genetic and molecular aberrations typically occurring in cancer. In this context, we will also present common characteristics of malignant tumors. Additionally, we will give a brief overview of current treatment options for cancer and present *personalized medicine* as a holistic approach to cancer prevention, diagnosis, and treatment.

2.1 Molecular biology of eukaryotic cells

In 1866, the Augustinian friar Gregor Mendel published the results of his archetypal experiments on plant hybridization and thereby provided, amongst others, first evidence for the fact that the heritable material of an organism is a significant determinant of the observed phenotype [16]. Nowadays, the elucidation of the relationship between the genotypic markup of an organism and its phenotypic manifestation is still a major focus of biology research. The 'flow of genetic information' is a conceptual scaffold that portrays this dependence as a multistep process in which information from the (epi-)genome is selectively expressed in the form of proteins and functional RNAs that accomplish various functions in the cell and that significantly contribute to the cell's phenotype (cf. **Figure 2.1 A**) [17].

The human body consists of 30 to 40 trillion individual cells of various specialization, distinct metabolism, and physiology [18]. The creation and maintenance of these different cellular phenotypes require the coordinated expression of particular sets of genes while others have to be repressed [19]. Numerous regulatory elements and mechanisms govern gene expression by impeding (or enabling) the flow of genetic information (cf. **Figure 2.1 B**).

The modulation of the 'packing' of genetic material in the cell's nucleus is a primary control point of gene expression. In the nucleus, the DNA is present in a highly condensed form, the chromatin. To achieve this high level of compression, the DNA is tightly wrapped around histones, which are octameric protein complexes, forming nucleosomes that are further folded into chromatin fibers [20]. Depending on the compression level of the chromatin, heterochromatin and euchromatin can be distinguished. Chromatin in its more densely compacted form is called heterochromatin. In heterochromatin, regulatory target sequences are likely to be inaccessible. Conversely, in euchromatin, the less compacted form of chromatin, these regulatory regions are reachable. In these open chromatin regions, transcription factors (TFs) can bind to specific control sequences in regulatory regions of a gene (e.g., the promoter, enhancer, or silencer regions), where they foster or repress the transcription of their respective target genes [21]. However, most transcription factors cannot exercise their function on their

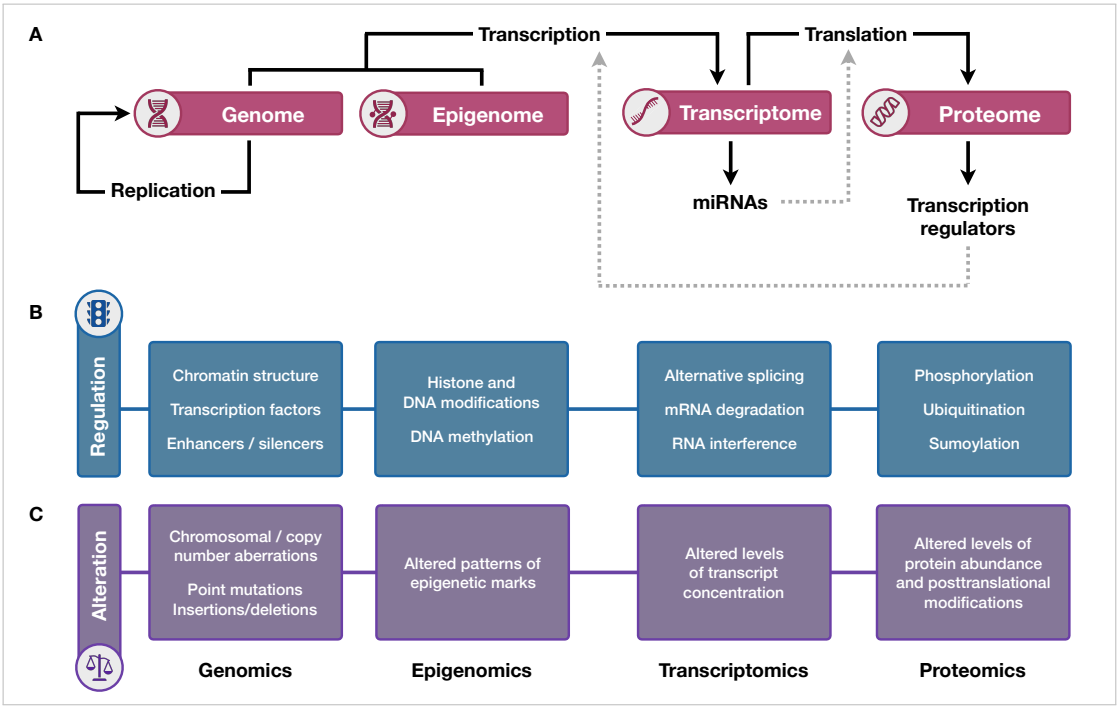


Figure 2.1 **Flow of genetic information**. **A)** Black arrows indicate the flow of genetic information from the genome to the proteome in a simplified scheme. The regulatory effects of miRNAs and transcription regulators are indicated by gray arrows. **B)** Examples of relevant regulatory elements and mechanisms controlling the flow of genetic information to the next level. **C)** Examples of measurable alterations that can help to elucidate the genetic and molecular underpinnings of a tumor's phenotype. The icons in this figure were obtained from [15].

own. Various other classes of proteins like coactivators or chromatin remodeling complexes are typically required to ensure the recruitment and activation of the transcriptional machinery, as well as proper elongation and controlled termination of transcription [22]. Besides the factors mentioned above, also epigenetic components play essential roles in governing gene expression. For example, there is a large variety of epigenetic marks, e.g. covalent modifications of histones or the DNA itself, which form complex regulatory patterns fostering either an activated, poised, or repressed state of the affected gene [23]. In general, acetylation of histone residues is associated with potentially active gene expression [24], while DNA methylation of CpG-rich regions close to the transcription start site of a gene is typically an indicator of gene silencing [25].

Once DNA transcription is successfully initiated, DNA can be transcribed into messenger RNA (mRNA) or non-coding RNA. In the case of mRNA, the primary transcript is called pre-mRNA, which can be made up of one or several exons and introns [26]. The pre-mRNA undergoes several post-transcriptional processing steps, which can involve 5'-capping, 3'-polyadenylation, alternative splicing, and sometimes RNA editing [27, 28]. The class of non-coding RNAs contains numerous types of functional RNAs that are not translated into proteins. These include, but are not limited to, ribosomal RNAs, transfer RNAs, and long non-coding RNAs [29]. Another prominent type of non-coding RNAs are microRNAs (miRNAs). These rather short nucleotide chains can exert their regulatory effect by causing the degradation of their target mRNAs or inhibition of their translation [30].

The mature mRNA, if not degraded or inhibited by regulatory RNA molecules, is transported to the ribosomes, where it is translated into the corresponding amino acid chain via complementary

base-pairing of tRNA adaptor molecules to the mRNA template [31]. The newly synthesized polypeptide then needs to fold into its three-dimensional structure, which ultimately defines the function of the protein. The process of protein folding can be supported by a complex cellular machinery of molecular chaperones. These chaperones support protein folding as well as the assembly of protein complexes [32]. In order to be able to react to external stimuli quickly, proteins can be activated or inactivated via post-translational modifications (PTMs) [33]. The most prominent type of PTM is phosphorylation, which describes the addition of a phosphate group to a protein, thereby quickly switching the protein's state from inactive to active or *vice versa*. Interacting proteins can form cascades of protein phosphorylations and dephosphorylations, transmitting and amplifying external signals (e.g., growth signals) into the nucleus. Besides signal transduction, PTMs can also determine a protein's cellular location or mark them for degradation. Examples of such PTMs are sumoylation and ubiquitination [34, 35]. Proteins, their interaction partners and substrates form complex and dynamic systems within cells. However, as interactions are not restricted to single cells, proteins can also receive signals from or convey signals to other cells [36]. In cases where this well-orchestrated interplay of gene regulation, protein-protein signaling, and the cellular metabolism gets out of balance, complex diseases may arise, a very prominent example of which is cancer [37].

2.2 Cancer

Over the last centuries, major advances in technology, sanitation, and medicine have significantly increased life expectancy in more and more regions of the world [38], shifting the leading causes of morbidity and mortality from infectious diseases to non-contagious diseases like cardiovascular diseases, diabetes, or cancer [39]. Cancer is a class of complex diseases that are characterized by the uncontrolled proliferation of previously healthy cells that gained the ability to invade adjacent tissues and to spread to distant parts of the body. In the following sections, we will describe the process of tumor development and a set of common cancer characteristics. Furthermore, we will review how these characteristics are exploited in various types of cancer treatment.

2.2.1 Cancer development and characteristics

The development of malignant neoplasms from healthy cells occurs in a multistep process during which the cells accumulate (epi-)genomic aberrations. During the lifespan of an organism, the DNA and its structure are continuously subject to impairment by various types of aberrations. DNA damage can be induced by erroneous DNA replication or environmental factors like chemical agents, radiation, or viruses [40, 41]. In healthy cells, damage detection and repair mechanisms counteract these aberrations either by repairing compromised DNA or labeling the cell for destruction. However, in cases where the cells' defense and control mechanisms are disturbed, aberrations might manifest in the genome [42]. In a Darwinian-like evolutionary process, those alterations that confer a growth advantage to the affected cells are selected for. Progressively, these cells accumulate further aberrations, potentially leading to the formation of a tumor [42].

2.2.1.1 (Epi-)genomic aberrations contributing to cancer

There are various types of genetic aberrations (i.e., mutations) that enable tumor pathogenesis (cf. **Figure 2.1 C**). Mutations can be classified with respect to their hereditability, size, and effect. Concerning hereditability, germline and somatic mutations can be distinguished. Changes occurring in an organism's germ cells (ova or spermatozoa) are called germline mutations. Germline mutations can be passed on to the offspring of an organism, which will carry these aberrations in all of its cells. There are several types of germline mutations known to confer a predisposition for specific cancer types, like breast cancer [43], colon cancer [44], and others [45]. Mutations arising in any other cell type (i.e., the 'soma') are called somatic mutations. The spread of somatic mutations is limited to the descendants of the affected cells. Although many cancers also have hereditary components, the occurrence of additional somatic mutations during the lifespan of an organism is typically required to initiate tumor pathogenesis [46].
Besides a classification by hereditability, mutations can also be classified by means of the size of the affected genomic region. Small genomic aberrations like point mutations or the insertion, deletion, or substitution of a few nucleotides can be distinguished from structural genomic alterations like copy number variations (i.e., amplifications and deletions) or the rearrangement of even larger genomic regions (e.g., resulting in fusion genes). Also on the level of whole chromosomes, structural and numeric anomalies can contribute to the development of cancer [47].
As a third way of classifying genomic aberrations, we can consider the effect of a particular mutation on the phenotype. Mutations can occur in coding regions of the genome, where they can affect the structure and hence the functionality of the encoded proteins, and in regulatory sites where they can contribute to alterations in gene expression. When considering protein-coding small-scale mutations, we can further classify mutations according to their effect on the protein structure. Here, we can distinguish synonymous and non-synonymous mutations. Synonymous mutations do not alter the encoded amino acid and hence do not affect the protein's structure and functionality. Non-synonymous mutations, on the other hand, can affect the encoded amino acid, the length of the polypeptide, or the reading frame. There are different types of non-synonymous mutations: Missense mutations result in the exchange of a single amino acid in the translated polypeptide. Nonsense mutations lead to the substitution of an amino acid with a premature stop codon. Readthrough mutations, on the other hand, induce the opposite, namely the replacement of a stop codon by a regular amino acid-encoding codon. The latter two types of non-synonymous mutations lead to truncation or elongation of the produced isoforms. In cases where an insertion or deletion ('indel') has a size that is not a multiple of three, this results in the shift of the open reading frame (frameshift mutation) that severely disturbs the peptide sequence.
Besides aberrations directly affecting the sequence and structure of the DNA, also epigenetic alterations ('epimutations') can contribute to carcinogenesis. Altered patterns of histone modifications and DNA methylation can affect the structure and integrity of the genome, as well as disrupting gene expression. A typical example of this is DNA hypomethylation of promoters, which can contribute to gene activation, whereas promoter hypermethylation can lead to the transcriptional silencing of the affected genes [48].
Tumors can potentially contain a plethora of different aberrations. However, not all of these contribute equally to a tumor's initiation and progression. Mutations that actually promote the

disease are referred to as driver mutations, whereas other 'bystander' mutations not central to the disease development are called passenger mutations. Whether an (epi-)genomic aberration is classified as driver or passenger mutation depends on various factors including the specific gene being affected as well as the mutation's effect on the activity of the corresponding protein. Here, two extreme cases can be distinguished: gain-of-function and loss-of-function mutations. Gain-of-function mutations can be achieved by various mechanisms like activating mutations, amplifications, or gene fusion to an actively transcribed gene. In the context of cancer initiation and progression, genes for which a gain-of-function mutation increases cancer risk are called proto-oncogenes (and 'oncogenes' in their mutated, hyperactive form). Typical classes of proteins that act as (proto-)oncogenes are transcription factors, chromatin remodelers, growth factors and their receptors, as well as signal transducers. Loss-of-function mutations can be mediated by inactivating mutations, deletions, or loss of chromosomal arms. Genes for which a loss-of-function mutation supports tumorigenesis are called tumor suppressor genes. Tumor suppressor genes are commonly involved in the negative regulation of cell proliferation.

2.2.1.2 The Hallmarks of Cancer

The (epi-)genomic alterations accumulated during tumorigenesis propagate to higher levels of cellular regulation, where they disturb signaling cascades and, ultimately, the behavior of the affected cells to enable malignant growth. Although the types and succession of the aberrations underpinning this transition can vary strongly between tumors, tumors can still be characterized by several commonalities. Hanahan and Weinberg summarized these common characteristics as the *Hallmarks of Cancer* [49]. The authors defined eight acquired capabilities and two enabling factors that allow cancer cells to survive, proliferate, and ultimately to metastasize (cf. **Figure 2.2**).

In brief, in order to allow malignant growth, cancer cells bypass the dependence of exogenous growth signals, which are required for healthy quiescent cells to proliferate again, by generating their own growth signals [50]. This mechanism is accompanied by cancer cells' insensitivity to anti-growth signals [51].

However, this growth potential would be restrained in healthy cells by the limited number of growth-and-division cycles cells are allowed to undergo before they reach a state of senescence or apoptosis. This limitation is attained by the progressive erosion of chromosome-protecting telomeres at each round of replication. Tumor cells counteract this process by the upregulation of telomerase expression [52].

Continuous cell growth and division need to be fuelled by large amounts of energy. To this end, cancer cells modify the cellular metabolism, for example via an increased uptake and utilization of glucose that, together with glutamine, is a major building block of cell maintenance and biosynthesis in mammalian cells. [53].

Another perspective on malignant growth is that cancer cells do not only grow and proliferate unlimitedly, but also feature the capability of evading apoptosis [54]. Resistance to apoptosis can be acquired by cancer cells via a multiplicity of mechanisms like suppression of mitochondrial membrane permeabilization and others [55].

Similar to the evasion of apoptosis via cell-intrinsic mechanisms, cancer cells also acquired the capability to avoid recognition and elimination by the immune system. In the early stages of

Figure 2.2 **The Hallmarks of Cancer**. Overview of a tumor's acquired characteristics and enabling factors (marked by an asterisk) promoting unlimited growth and disease progression. Figure based on [49]. The icons in this figure were obtained from [15].

tumorigenesis, the immune system typically can identify and destroy incipient cancer cells. However, in a process called immunoediting, weakly immunogenic cancer cells are selected for. The reduction in immunogenicity can be mediated by various mechanisms like defects in or a total loss of target antigen presentation [56]. Even in cases where the recognition of tumor cells is still functioning, cancer cells can evade immune destruction by the expression of immunoregulatory checkpoint proteins like PD-L1 that bind to corresponding receptors on activated T-cells, thereby preventing them from eradicating the malignant cells [57].

In order to supply the increasing amount of tumor tissue with oxygen and nutrients, the cancer cells induce the sprouting of new blood vessels from the existing vascular system, i.e. angiogenesis [58].

However, lack of further space and nutrients lets cancer cells develop the capability of invasion and metastasis, founding new colonies of malignant cells in distinct areas of the body, disconnected from the primary tumor [59].

Besides the eight hallmark capabilities described above, Hanahan and Weinberg define two enabling characteristics that foster their acquisition: genome instability and mutation as well as tumor-promoting inflammation. In order to transform healthy cells into cancer cells, checkpoints preventing cells with DNA mutations from proliferation, have to be circumvented. As a consequence, the mutability of the cell's genome is increased, which enables the acquisition of the described features [60]. Inflammation can foster multiple hallmark capabilities, for example by supplying growth factors and extracellular matrix-modifying enzymes to the tumor microenvironment that can promote typical hallmark characteristics like angiogenesis, invasion, and metastasis [61].

2.2.2 Cancer treatment

Typical treatment options for tumors include surgery, radiation, systemic chemotherapy, and in the recent past also immunotherapy. Depending on a variety of factors, such as tumor size and stage, its location, and the overall health status of the patient, different treatment modalities or combinations thereof are utilized. The surgical excision of a tumor is the primary treatment option for many cancer types, especially when diagnosed at an early stage [62]. However, the applicability of surgery is limited when tumors are difficult to reach or have metastasized.

Another common treatment option for solid cancers is radiotherapy. Here, a beam of ionizing radiation is targeted at the tumor. The absorbed energy leads to the formation of free radicals in the cancer cells, which severely damage the DNA and ultimately cause cell death [63].

For systemic therapy, there exists a plethora of drugs: At the time of writing, the National Cancer Institute (NCI) lists more than 250 approved anticancer drugs and combination regimen [64]. The approved drugs can be subdivided into two major categories: non-specific chemotherapy and targeted therapy. Conventional non-targeted cytotoxic agents encompass several classes like taxanes, anthracyclines, or alkylating agents [65]. Although these drug classes have different mechanisms of action, they all rigorously attack rapidly dividing cells in the body. This high promiscuity leads to severe side effects as for example also cells of the gastrointestinal epithelium and the immune system are damaged.

In contrast to this, targeted drugs try to utilize characteristics of the tumor to specifically attack cancer cells, thereby maximizing efficacy while minimizing toxicity. The rationale behind this

approach is a concept called 'oncogene addiction', which postulates that tumors are likely to be dependent on a single or a few (potentially druggable) cancer genes for the manifestation of their malignant phenotypes [66].

With respect to targeted chemotherapy, we can distinguish two main classes of therapeutic agents: small molecules and monoclonal antibodies [67]. Small molecules are organic compounds that are able to enter cells due to their low molecular weight. Small molecules in cancer treatment are typically tyrosine or serine/threonine kinase inhibitors (indicated by the suffix '-ib'), which target aberrant signal transduction processes in the cell [68]. Examples of such kinase inhibitors are imatinib for the treatment of chronic myelogenous leukemia via inhibition of the BCR-ABL tyrosine kinase [69], or vemurafenib, which targets BRAF-mutant advanced melanoma [70]. Besides small molecules, monoclonal antibodies (indicated by the suffix '-mab') are an emerging and highly specific means of cancer treatment. Antibodies can occupy cell surface receptors and thereby block essential signaling cascades responsible for the propagation of, for example, growth signals into the nucleus. Additionally, they can trigger antibody-dependent cellular cytotoxicity. A prominent example of such a highly specialized drug is trastuzumab, a humanized monoclonal antibody targeting the human epidermal growth factor receptor 2 (HER2), which is over-abundant in some types of breast cancer [71]. Other examples are bevacizumab, which targets angiogenic processes in the tumor by inhibiting the vascular endothelial growth factor A (VEGFA) [72] and rituximab for the treatment of diffuse large B-cell lymphoma [73]. In recent years, antibody-drug conjugates (ADCs) have risen to become a promising class of targeted cancer treatment. ADCs combine the high selectivity of monoclonal antibodies with the cytotoxic activity of traditional chemotherapeutic drugs. To this end, small molecules with high toxicity are covalently bound to an antibody via a linker that is stable in circulation but releases the cytotoxic agent when bound to the target cell [74].

Targeted drugs have significantly advanced cancer treatment over the last two decades. They were commonly considered as 'magic bullets', a term originally coined by the bacteriologist Paul Ehrlich in referral to chemicals specifically targeting microorganisms [75]. However, the individual administration of highly specialized compounds does not properly account for the multidimensional nature of cancer, which can involve many molecular players interacting in interconnected pathways. Hence, the treatment of a tumor with a single targeted drug, aiming at the ablation of a single molecular target or pathway is unlikely to achieve complete remission. Under the selective pressure of the drug, resistant subpopulations emerge that can quickly outgrow the sensitive cells and cause relapse [76]. There are various mechanisms by which tumors evade treatment, for example by mutating and structurally altering drug targets such that inhibitors cannot bind anymore or via the activation of alternative signaling pathways to alleviate the dependence on signal transduction cascades blocked by a therapeutic agent [77]. To prevent these escape mechanisms, combination therapies accounting for several mechanisms driving a tumor have proven great potential [78–80]. Combination schemes can follow various strategies like the maximal inhibition of a single target (e.g., bevacizumab-sunitinib against VEGF and its receptor VEGFR), targeting multiple key players 'vertically' along a pathway (e.g., HER2 and mTOR in the ErbB2 signaling pathway), or attacking several 'parallel' oncogenic signaling pathways (e.g., the VEGF signaling pathway plus the EGFR signaling pathway) [81]. Besides the presented approaches to cancer treatment, cancer immunotherapy has been gaining more and more attention as a cancer treatment option, especially in the last years. The

term immunotherapy encompasses various approaches that all aim at harnessing the patient's immune system to treat cancer. Emerging immunotherapies for cancer include checkpoint blockade, adoptive T-cell therapy, and therapeutic cancer vaccines.

Typically, the recognition of tumor antigens by effector T-cells leads to the destruction of the tumor cells. However, as described in **Section 2.2.1**, cancer cells are likely to evade detection and eradication by the immune system. One central mechanism by which this can be achieved is the activation of immune system-suppressing checkpoints like CTLA-4 or PD-1 [82]. For instance, in cases where cancer cells express the programmed death ligand 1 or 2 (PD-L1/2) on their surface, the PD-1 receptor on T-cells in close proximity may bind to this ligand inducing an inhibitory signal and preventing the tumor cell from being destroyed. In such cases, inhibiting either the receptor or the ligand (e.g., by pembrolizumab binding to PD-1 [83] or atezolizumab against PD-L1 [84]) can allow T-cells to resume their attack of the tumor cells. Several studies across cancer types showed that the effectiveness of such checkpoint inhibitors - amongst others - correlates with the number of mutations present in a tumor, making them a promising treatment option, especially for advanced and aggressive tumors [85, 86].

Besides checkpoint blockade, adoptive T-cell therapy is another promising approach to cancer immunotherapy [87]. In order to boost a patient's immune system, T-cells are harvested from the patient, grown *in vitro* and reintroduced in considerable amounts. There are also forms in which the T-cells are genetically modified or chimeric antigen receptors are attached to the T-cells, such that they can better recognize cancer cells.

Another approach to induce the formation of tumor-specific cytotoxic T-cells is the use of personalized cancer vaccines [88, 89]. Cancer vaccines aim at overexpressed or altered proteins and HLA-presented peptide sequences (neoantigens) that resulted from genetic, epigenetic, and gene expression changes uniquely characterizing the patient's tumor. They are used to prime T-cells to recognize these characterizing antigens and induce a T-cell mediated immune reaction. As the neoepitopes are dependent on both the patient's tumor mutations and HLA genotype, cancer vaccines have to be individually designed. The identified epitopes can serve as the basis for the synthesis of a personalized cancer vaccine, which can be combined with checkpoint inhibitors to potentially boost the effectiveness of the vaccine [90].

Although the diagnosis and treatment of cancer have greatly improved over the last decades and survival rates across cancer types have generally increased, mortality rates and therapeutic success still vary strongly between cancer types (cf. **Figure 2.3**). Besides the depicted disparity between different types of cancer, there is also a high genotypic diversity between tumors of the same type, subtype, and even between different regions of the same tumor that impede therapy stratification and hence must be accounted for in the treatment decision-making process [91].

2.3 Personalized medicine

Personalized medicine, often also referred to as precision medicine or individualized medicine, is a holistic approach of tailoring disease prevention, diagnosis, treatment, and monitoring to an individual based on his/her genetic and molecular makeup, clinical data, and medical history [93]. Although physicians have always strived to find the optimal treatment for their patients, the emergence of high-throughput biotechnological techniques since the turn of the last millennium marks an inflection point in the diagnosis and treatment of complex diseases.

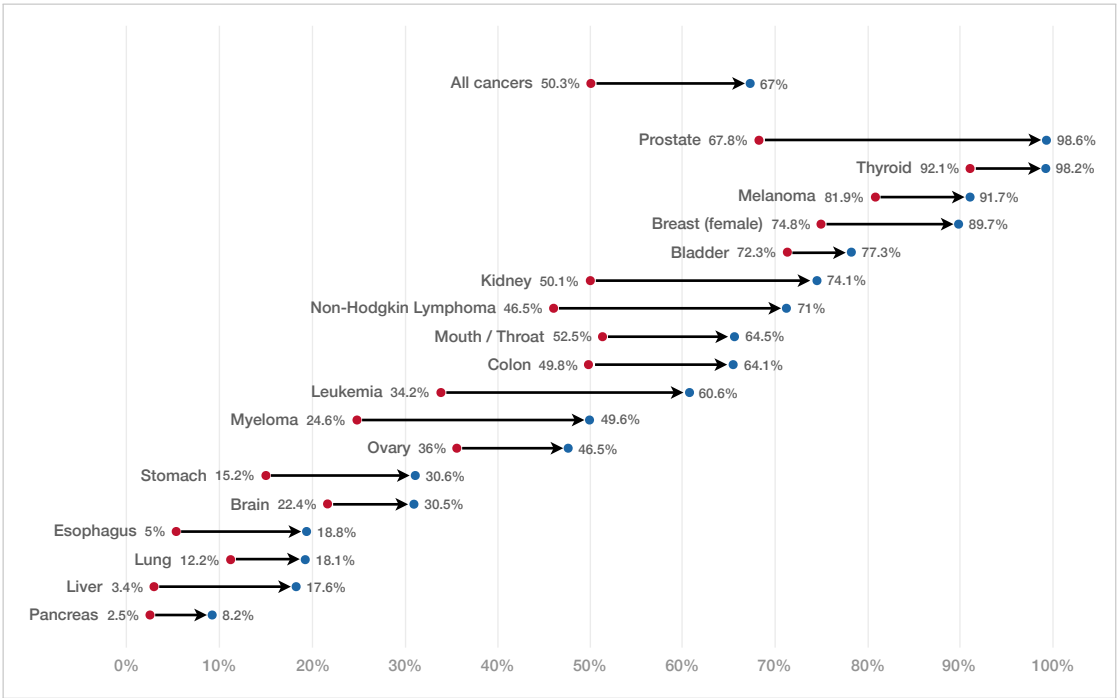


Figure 2.3 **Five-year cancer survival rates in the USA**. Average five-year survival rates for various cancer types in the United States. Red points indicate the rates for 1970-77 and the blue points indicate the rates for 2007-2013. Data obtained from [92].

DNA and RNA sequencing [94], microarray technologies [95], proteomics approaches [96], and other techniques provide an unprecedented wealth of data that can be used for individualized phenotyping.

The concept of personalized medicine is used in a broad set of diseases, including infectious diseases, cardiovascular disorders, or neurological conditions [97]. One of its major fields of application is cancer, for which various aspects of personalized intervention will be exemplified in the following.

One of the goals of personalized medicine is to fulfill a paradigm shift from reactive to preventive healthcare [98]. Predisposition testing is thereby a major cornerstone to identify individuals at risk for certain diseases. For example, women carrying harmful germline mutations in the tumor suppressor genes BRCA1 or BRCA2 have an up to eight-fold increased risk to develop breast cancer and an up to 40-fold risk to develop ovarian cancer during their lifetimes [99]. For patients at increased risk, screening frequency could be increased to detect developing tumors as early as possible [100], they could undergo a prophylactic mastectomy in the case of breast cancer [101] or be treated with chemopreventive agents [102].

In cases where the occurrence of a malignancy could not be prevented, there are various options for the personalized management of the disease. Considering radiotherapy, personalization could be achieved via an adaptation of the radiation dose to the tumor's specific composition. For example, a PET/MRI scan could be used to identify aggressive hypoxic regions in the tumor that require a higher intensity of radiation in these areas [103]. In the realm of systemic therapy, there is also a large potential for personalization. Numerous factors need to be taken into account to determine the potential efficacy of a drug or drug regimen for a given tumor. Personalized cancer chemotherapy aims at targeting the pathophysiologically and therapeutically relevant

alterations that drive a specific tumor. In order to identify potential treatment options, these tumor characteristics have to be identified and assessed with respect to their effect on drug sensitivity. In addition to tumor-specific mutations, also aberrations affecting genes involved in absorption, distribution, metabolism, and excretion (ADME) of drugs are of great importance for treatment selection and dosing. The assessment of drug-processing enzymes like members of the cytochrome P450 family or transporter proteins is also referred to as pharmacogenomic (PGx) testing. Genomic aberrations in cytochromes can affect the metabolism of a wide range of commonly used drugs. They can affect the rate at which drugs are metabolized in the body or even completely prevent the activation of prodrugs in the liver [104].

Across all stages of personalized medicine, but specifically prior to and during treatment, biological markers (biomarkers) play an essential role in pinpointing relevant tumor characteristics. As defined by the Food and Drug Administration (FDA), a biomarker is a "characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions" [105]. Biomarkers can be categorized according to their purpose into several classes, including susceptibility biomarkers, diagnostic biomarkers, prognostic biomarkers, predictive biomarkers, and monitoring biomarkers [105]. For treatment decision support, prognostic and predictive biomarkers are of major interest. Prognostic biomarkers provide information on the progression of the disease. Prominent examples of prognostic biomarkers are Oncotype DX [106] and MammaPrint [107] that determine the aggressiveness and risk of relapse of a breast tumor based on multi-gene signatures, thereby also assessing the need for adjuvant chemotherapy. In contrast to prognostic biomarkers, predictive biomarkers are concerned with the efficacy and safety of certain drugs or types of treatment. *In vitro* diagnostic tests or imaging tools to assess the state of predictive biomarkers are called companion diagnostics (CDx) [108].

Nowadays, there are more than 40 types of CDx approved for cancer treatment stratification by the FDA. These include the interrogation of HER2 amplification status prior to treatment with trastuzumab or pertuzumab in breast cancer [109], the assessment of KRAS mutation status for cetuximab treatment in colon cancer [110], or the recognition of the Philadelphia chromosome (BCR-ABL fusion gene) in chronic myeloid leukemia to ensure susceptibility for imatinib [111] or nilotinib [112].

Moreover, companion diagnostics and biomarker-based treatment stratification can also be utilized in the drug development process. Besides rational drug design, the concept of drug repositioning plays a major role in this context. Drug repositioning describes the identification of new indications for already approved drugs [113] or drug candidates that failed to reach market entry due to various reasons (the latter sometimes referred to as 'drug repurposing') [114]. The lung cancer drug gefitinib, for example, could not demonstrate a survival advantage in the general population of lung cancer patients and was withdrawn from the market after initial FDA-approval. However, in the subgroup of patients with specific EGFR mutations, gefitinib could show significant benefits and hence was approved as first-line treatment for accordingly stratified lung cancer patients in 2015 [115]. Also, the expansion of indications for a drug can be observed: Crizotinib, a kinase inhibitor approved to treat specific forms of non-small cell lung cancer, including those that are EML4-ALK positive, is also effective in other tumors containing ALK alterations, such as anaplastic large cell lymphoma [116] or pediatric neuroblastoma [117]. In 2017, the FDA even granted tissue/site-agnostic approval to

the checkpoint inhibitor pembrolizumab, which was the first targeted cancer treatment to be administered across cancer types for patients with metastatic, microsatellite instability-high or mismatch repair-deficient solid tumors [118].

With an increasing body of knowledge about a tumor's underlying genomic alterations and the expression of relevant biomarkers, tumor classification and patient stratification are shifting away from the tissue of origin and toward a molecular taxonomy and mechanism-driven treatment decisions. This paradigm shift is also reflected in the way clinical trials are conducted. Traditionally, clinical trials were mostly focused on average responses across a population, not on the specific subgroups that might be especially susceptible to the treatment due to their (epi-)genetic and molecular markup. Nowadays, there are new and emerging classes of precision medicine clinical trials that account for personalization: so-called 'basket' and 'umbrella' trials. In basket (or bucket) trials, a drug of interest targeting a specific aberration is tested on a variety of tumor types in a biomarker-positive subgroup. Umbrella trials have many different treatment arms (like the spokes of an umbrella) in which patients are matched to different drugs depending on the molecular makeup of their disease. The recently completed prospective umbrella trial MOSCATO 01 evaluated the clinical benefit of high-throughput genomic analyses guiding treatment selection in advanced solid tumors. However, only for 411 of 843 samples an actionable genomic alteration could be identified and of the accordingly treated patients only 11% showed a response to their matched treatments [119]. Also, the currently running NCI MATCH trial, the largest combined basket/umbrella trial to date, could so far only assign a genetics-based treatment to 9% of the enrolled patients [120].

Hence, in order to comprehensively characterize and stratify a given tumor, besides of the sole consideration of individual actionable mutations, the complex molecular 'circuitry' of altered genes and proteins and their interdependencies also need to be taken into account to obtain a holistic view on disease-driving mechanisms [121].

3

Materials and Methods

A general goal of life science research is the identification of the components that make up a living system. Specifically, the elucidation of the relations and interactions among these components that result in the functioning of the system as well as its dysfunctioning in the presence of perturbations, are of major interest.

With respect to the personalized treatment of cancer, this goal translates to (i) the identification of (epi-)genomic and molecular aberrations present in a tumor and (ii) the evaluation of these alterations concerning disease initiation and progression as well as potential treatment responses [122]. In this chapter, we will give an overview of the biotechnological techniques, computational and statistical methods, as well as reference databases that are required to approach these goals. To this end, we will first introduce three prevailing types of high-throughput experimental techniques and corresponding bioinformatics processing steps to comprehensively characterize tumor samples with respect to their (epi-)genetic and molecular profiles. In order to extract relevant information from these high-dimensional and complex data sets, *a priori* knowledge from various domains and robust statistical methods are required. Hence, in the second part of this chapter, we will present several classes of databases and resources that cover a variety of pathway-, disease-, and treatment-specific aspects including the functional annotation of genes and gene sets, known cancer driver genes, drug-target interactions, and the pharmacogenomic effects of mutations. Finally, we will present various types of statistical tests and algorithms for the identification of deregulated genes and biological processes.

3.1 High-throughput experimental techniques

The totality of specific types of biological entities (e.g., genetic material, transcripts, or proteins) is typically indicated by the common suffix '-ome', as for example in 'genome', 'transcriptome', or 'proteome'. Current biotechnological high-throughput techniques allow to comprehensively study different types of 'omes' and hence are referred to as 'omics' technologies [123]. The thereby obtained high-dimensional data sets ('omics data') can be used to address various types of research questions ranging from basic science to translational research and clinical decision support. In the following sections, we will describe three main classes of biological assays that can be used to measure (epi-)genomics, transcriptomics, and proteomics data. For each of these techniques, the experimental principle, areas of application, and required computational steps for data processing will be presented.

3.1.1 Microarrays

The microarray technique is based on the highly parallel hybridization of entities under investigation (e.g., DNA, RNA, or proteins) to a predefined set of specific, complementary probes on a slide, the actual microarray. Fluorescence intensities are captured to assess the degree of hybridization, which acts as a proxy for the quantification of the entity under investigation [95].
In the following, the experimental principle of microarrays will be presented using the example of cDNA microarrays and one of its predominant applications: the measurement of transcript levels. Afterward, further types of microarrays to measure various other kinds of cellular aberrations will be addressed. Finally, the required processing steps for the quantification of the measurements will be discussed.

3.1.1.1 Experimental principle

A typical application of cDNA microarrays is the assessment of gene expression at the mRNA level. An overview of the workflow is provided in **Figure 3.1**.

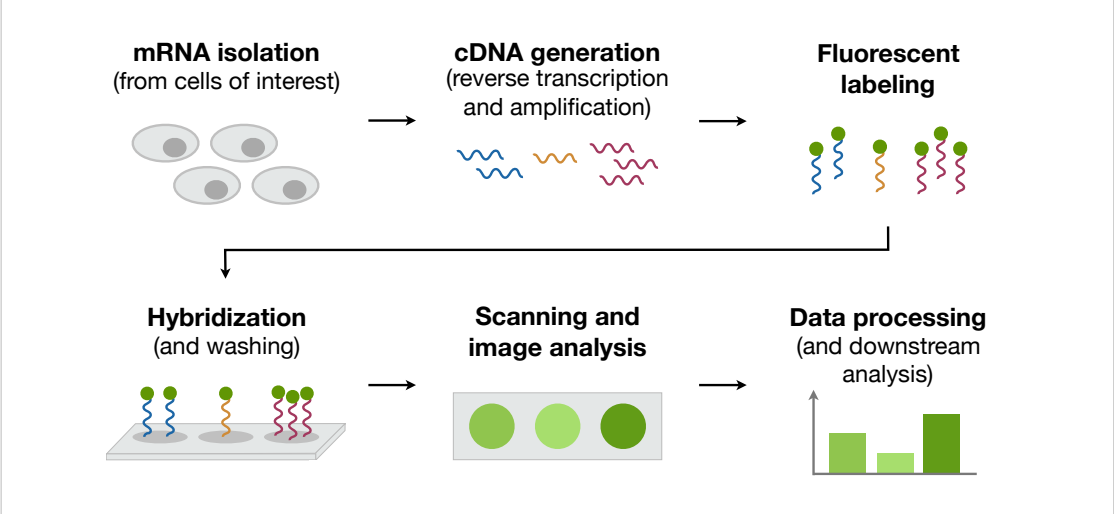


Figure 3.1 **Overview of cDNA single-channel microarray workflow**. After extraction of mRNA from the sample of interest, the mRNA is reversely transcribed into cDNA and amplified. Fluorescent labeling of the probes is followed by hybridization to the microarray and a washing step. (Please note that the microarray probes are not depicted for improved visual clarity.) Based on the intensity of the emitted fluorescence in the respective spots, expression levels of transcripts and genes can be deduced. Figure adapted from [124].

In a first step, the mRNA is extracted from the cells under investigation, which is reverse-transcribed and amplified to generate complementary DNA (cDNA) for hybridization. The cDNA is afterward labeled with a detectable marker, typically a fluorescent dye like the fluorophore Cyanine 3 (Cy3) [125]. The microarray itself is a small platform consisting of a solid surface on which specific oligonucleotide probes are bound to designated spots. The fluorescence-labeled cDNA is hybridized onto the microarray, relying on the high specificity of complementary base-pairing [126]. The hybridization is followed by a washing step that aims at removing unbound as well as unspecifically binding cDNA. The degree of hybridization for each spot is assayed by measuring the emitted fluorescence using, e.g., a confocal laser scanner.

Several steps of subsequent image analysis and data normalization (cf. **Section 3.1.1.3**) then yield expression levels that can be compared to those of a reference sample or control group to identify sample-specific expression changes (cf. **Section 3.3.2**).
Besides the single-channel experiments described above, there are also two-channel experiments in which two samples to be compared against each other are labeled with fluorophores emitting at different wavelengths (e.g., Cy3 and Cy5) and compete for hybridization to the microarray probes [127]. In this setting, differential expression can be directly deduced from the relative fluorescent intensities.

3.1.1.2 Other microarray technologies and applications

There are also numerous other experimental techniques that use different types of microarrays, three of which will be described in the following.

Comparative genomic hybridization microarrays: Besides the assessment of transcript levels, also copy number variations between samples (cf. **Section 2.2.1.1**) can be assessed using microarrays, in an experiment type called Comparative Genomic Hybridization on arrays (aCGH). Here, the microarray contains probes of single-stranded DNA fragments of known chromosomal location. In a two-channel microarray setting, the DNA extracted from a test sample and a reference sample are denatured, labeled with fluorescent dyes of different colors, and applied to the microarray for competitive hybridization. The captured relative fluorescence intensities are used to compute relative copy numbers for the two samples. Depending on the research question at hand, probe lengths can range from less than 100 bases to several hundred kilobases. By this, the aCGH approach overcomes limitations of previous cytogenetic techniques like conventional Comparative Genomic Hybridization (CGH) [128], which only provided low resolution, and Fluorescence *In Situ* Hybridization (FISH) [129], which could only be used for the analysis of a limited number of chromosomal loci at a time [130].

Methylation microarrays: Microarrays can also be used to measure DNA methylation [131]. As described in **Section 2.1**, DNA methylation is a type of epigenetic modification, which is not readily detectable from the underlying genomic sequence. Hence, in order to be assessable using microarrays or sequencing technologies (cf. **Section 3.1.2**), methylation states have to be 'translated' to the nucleotide level. Bisulfite conversion is the current gold standard for the quantification of DNA methylation [132]. The treatment of the DNA under investigation with sodium bisulfite leads to the conversion of unmethylated cytosines to uracil, while methylated cytosines remain unaltered [133]. After fragmentation, amplification, and denaturation, the bisulfite-converted DNA is hybridized to the microarray. Here, commonly used technologies such as the Illumina Infinium HumanMethylation450 BeadChip [134] measure methylation levels at single CpG resolution using two types of probes, one complementary to the methylated allele and one to the unmethylated allele. The state of a specific CpG thereby is assumed to also be representative for flanking CpG sites as for example present in CpG islands. After the allele-specific annealing of the fragments, the probe sequence is extended by labeled nucleotides. This extension only occurs for those sequences that perfectly match the probes, i.e. for unmethylated loci only those that are bound to the probes for the unmethylated state and for methylated loci only those that are bound to the probes for the methylated state. The respective signal

intensities obtained from subsequent immunohistochemical staining are used to quantify DNA methylation levels as β-values (or M-values).

Protein microarrays: Another type of cellular characteristic measurable using microarrays is protein abundance. Techniques like Forward-Phase Protein MicroArrays (FPPAs) can be used to quantify protein levels [135]. In contrast to the types of microarray experiments described above, instead of oligomeric nucleotides, protein-specific antibodies are attached to the microarray. After protein extraction from the cells under investigation, the lysate is added to the microarray. The amount of protein binding to the probe antibodies can be quantified by using fluorescence labeling either of the lysate proteins themselves or via reporter antibodies in a so-called 'sandwich' assay format [136, 137]. Conversely, Reverse-Phase Protein MicroArrays (RPPAs) can be used to assess the abundance of a single type of protein across a large number of samples, which are individually spotted onto an array. Here, the levels of the protein under investigation are queried using a highly specific antibody [138]. In both cases, the reliability of the measurements is crucially dependent on the specificity of the employed primary (and potentially secondary) antibodies.
Moreover, there are also microarray platforms that use so-called 'aptamers' instead of antibodies. Aptamers are single-stranded oligonucleotide molecules that specifically bind to their target proteins. The SOMAscan assay by SomaLogic employs their proprietary SOMAmer (Slow Off-rate Modified Aptamer) reagents to capture up to 5,000 proteins simultaneously [139, 140].
While protein microarrays are an emerging means of protein identification and quantification [141], mass spectrometry-based approaches, which will be presented in **Section 3.1.3**, are another commonly used approach to proteomics.

3.1.1.3 Processing of microarray data

For obtaining expression values from microarray experiments, multiple processing steps are necessary, which will be exemplarily outlined for single-channel oligonucleotide microarrays (cf. **Section 3.1.1.1**) in the following paragraphs.

Scanning and image analysis: After the excitation of the hybridized probes on the microarray using a laser beam of fluorophore-specific wavelength, a gray-scale image of the microarray is captured by the laser scanner. In order to obtain the raw fluorescence signals for each entity on the array, several steps of image analysis are required. As microarray spots typically follow a grid-like arrangement, the pixels belonging to each probe can be identified by placing a grid onto the image, which is adjusted by geometric operations on the pixel rows and columns to account for slight offsets and rotations [142]. Then, the pixels in each target area are classified as belonging either to the signal ('foreground') or the surrounding area ('background'). This segmentation can be achieved using different approaches, including adaptive shape detection [143] or clustering methods [144]. Based on this classification and the intensities of the respective pixels, the overall intensity of a spot can be inferred.

Nonspecific-binding correction and summarization: Although the washing step in microarray protocols aims at removing unspecifically binding fragments, a residual amount of unspecific

hybridization typically remains, which has to be taken into account to identify and quantify expressed genes [145].
An example of how to account for this already by design of the microarray are Affymetrix oligonucleotide arrays. Here, each target sequence to be measured is represented by a probe set of up to $n = 20$ pairs of relatively short (25 nucleotides long) perfect-match (PM) and adjacent mismatch (MM) probes, which are distributed across the chip. PM probes are perfectly complementary to the targeted sequence, whereas MM probes contain a single-base substitution at a central position. The amount of hybridization to the MM probes is assumed to be representative of non-specific hybridization events. For each target sequence, the corresponding spots yield two series of values, $PM_i, ..., PM_n$ and $MM_1, ..., MM_n$. For a qualitative assessment of whether or not an actual signal can be detected for a given target sequence, the two series can be tested for a significant difference in their distribution (e.g., using the Wilcoxon signed-rank test, see **Section 3.3.2**). For the quantitative determination of expression of a target sequence s, the average difference between PM and MM partners can be considered:

$$\bar{s} = \frac{\sum_i^{n_s} (PM_i^s - MM_i^s)}{n_s},$$

where PM_i^s is the PM value of the i-th probe of target sequence s, MM_i^s is its corresponding MM value, and n_s is the number of probe pairs available for sequence s. Alternatively, there are also averaging-methods that account for potential outliers like the One-Step Tukey Biweight Algorithm [146].

Normalization: Besides unspecific hybridization, there are numerous other factors that may add noise or systematic sources of variability to the raw microarray data, for example, RNA degradation or a varying spotting efficiency [147, 148]. Hence, in order to ensure intra- and inter-array comparability, a normalization step is required. The background correction step described above can be considered as an intra-array normalization [149]. For inter-array normalization, there are various approaches, including the use of reference gene sets or distribution-matching procedures. The selection of a reference set of probes as normalization sample thereby relies on the assumption that these probes do not show much biological variation across experiments (e.g., via the use of spike-in mRNA) [150]. Distribution-matching procedures aim at making expression value distributions comparable across arrays. To this end, quantile normalization is a commonly applied method, which adjusts the expression value distribution across arrays [151]. An alternative to this is Variance Stabilizing Normalization (VSN), which is based on the observation that fluorescence measurements of high intensities typically show larger variances than low-intensity measurements. VSN transforms the data such that the variance is stabilized across the whole range of expression [152].
There are several algorithms and software packages available that aim at making expression value distributions comparable across arrays, of which Robust Multiarray Average (RMA) [153] and Affymetrix MicroArray Suite 5.0 (MAS 5.0) [154] are two of the most popular ones.

Batch effects: Microarray experiments do not measure gene expression in absolute units and they are highly sensitive to changes in the experimental setup as well as the external conditions under which the experiment is performed. As a consequence, the results of microarray

experiments conducted in different experimental runs will exhibit differences that are due to non-biological, but rather technical variance. These systematic errors are called batch effects and they are likely to impede subsequent data analysis steps if not accounted for [155]. The causes of batch effects can range from major differences in the experimental setup like the use of different types of microarray platforms or experimental protocols over variations in sample preparation to presumably minor factors like the temperature and lighting or the actual 'batches' of chemical reagents used [156, 157]. Batch effects arise whenever a set of samples is measured in multiple parts or microarray data sets from different resources are being combined. Hence, in these cases it is necessary to correct for the bias introduced by batch effects. To this end, batch effect removal techniques such as SVA [158], ComBat [159], or RUV-2 [160] should be applied. However, all of these methods require an experimental design that accounts for batch effect removal, where one or more samples of every condition should be contained in each batch. If this is not the case, the impact of the phenotype can hardly be differentiated from the impact of the confounding factor(s), making the reconstruction of the actual signal nearly impossible [161].

3.1.2 High-throughput sequencing

The term 'sequencing' generally describes the process of identifying the primary structure (the 'sequence') of larger molecules of covalently bound monomeric units, forming so-called 'biopolymers'. While this broad definition also covers polypeptides and polysaccharides, we will focus on polynucleotides (DNA and RNA) in the following. Nowadays three generations of DNA/RNA sequencing technologies are distinguished that differ in their throughput and experimental overhead: they range from first-generation and rather low-throughput Sanger sequencing [162] over the massively parallel sequencing of shorter fragments in second-generation technologies [163] to the 'real-time' sequencing of very long molecules in the third generation [164]. The term 'high-throughput sequencing' comprises the second and third generations of sequencing techniques.
There is a broad variety of protocols and techniques to conduct high-throughput sequencing that vary depending on the sequencing platform used and the task at hand. In the following sections, we will describe the underlying experimental principles, different areas of application, and the computational steps required for the processing of the experimental data.

3.1.2.1 Experimental principle

There are three main approaches to sequencing: (i) sequencing-by-synthesis, (ii) sequencing-by-ligation, and (iii) sequencing-by-hybridization [165]. While sequencing-by-synthesis uses polymerases to successively add individual bases to a growing strand [166], short oligonucleotides are used in the other two cases. Specifically, the formation of perfectly matching duplexes with a target sequence is considered in sequencing-by-hybridization [167]. In sequencing-by-ligation, hybridizing oligonucleotides are ligated to complement the template strand [94]. Besides different ways of base or oligonucleotide integration, also the means by which an incorporated base or bound fragment is identified differs between sequencing techniques. Many sequencing technologies rely on the optical detection of fluorescent reporters [168]. However, there are also approaches that rely on the detection of hydrogen ions that are released during polymerization (e.g., IonTorrent's semiconductor technology) [169] or the

change of electrical current as DNA bases pass through a nanopore (as in Oxford Nanopore's sequencing technology) [170].
In the following, we will explain the experimental principle of DNA sequencing via the example of the commonly used second-generation technique of sequencing-by-synthesis, as provided by Illumina for Whole Genome Sequencing [171]. **Figure 3.2** provides an overview of the workflow.

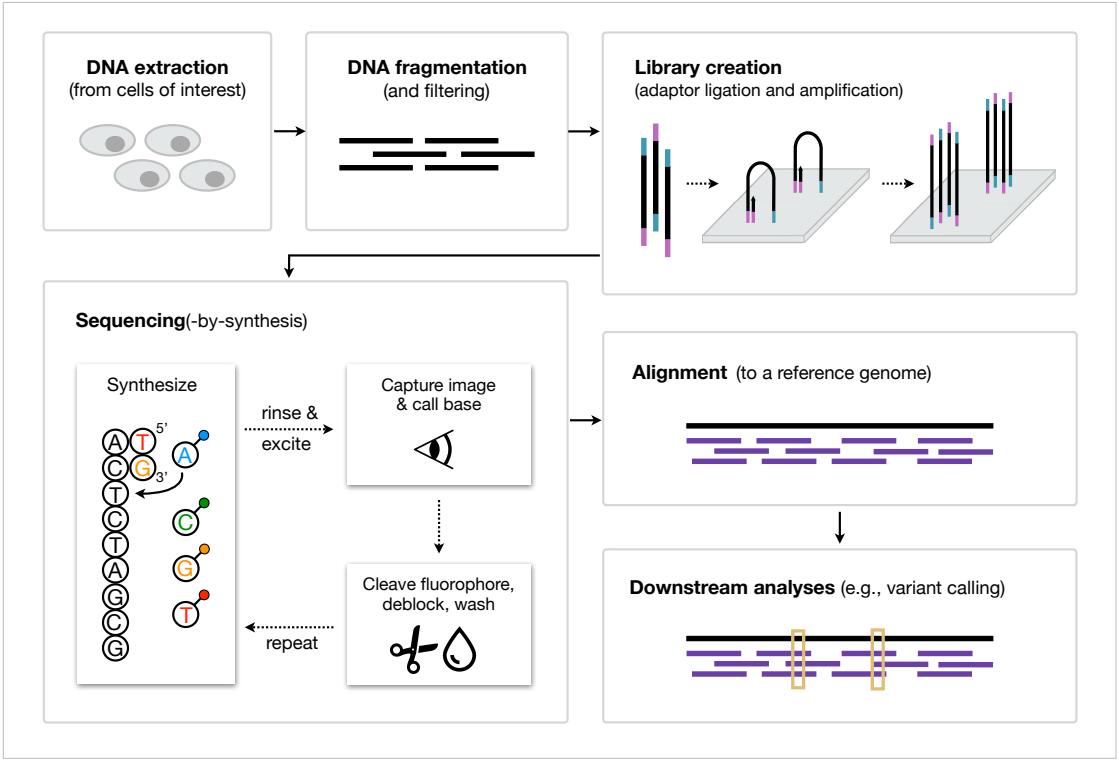


Figure 3.2 Overview of Illumina sequencing-by-synthesis workflow. DNA extraction and fragmentation are followed by the construction of the sequencing library. In this step, adapters are attached to the fragments, which are amplified on the flow cell using Bridge PCR to form clusters. The actual sequencing is performed by iterative rounds of nucleotide incorporation and fluorescence capturing. The thereby obtained sequencing reads can be aligned to a reference genome and downstream analyses like variant identification can be performed. Parts of the image adapted from [171] and [172]. The icons in this figure were obtained from [15].

In a first step, the DNA is extracted from the cells under investigation. Next, the analyte is fragmented using random shearing (e.g., via nebulization, isothermal sonication, or enzymatic fragmentation [173]) and filtered for fragments of protocol-appropriate length, typically between 200 and 500 bp [174]. The template is subsequently amplified using a specialized form of Polymerase Chain Reaction (PCR), called Bridge PCR. In this amplification technique, template DNA is denatured and two types of adapters are ligated to the respective ends of the fragments. These fragments are hybridized to the so-called 'flow cell' that contains a 'lawn' of adapter-complementary primers. After hybridization of the fragment adapters to the first type of primer, a polymerase creates the complement of the hybridized fragment. The double-stranded molecule is then denatured and the original template is washed away. The remaining strand bends over and hybridizes - in resemblance of a bridge - to the second type of primer and strand complementation is initiated. In the following PCR process, the double-stranded molecules are iteratively denatured, each strand forming a new 'bridge' for DNA replication

and thereby exponentially amplifying the template sequence. After clonal amplification of the fragments, the reverse strands are cleaved and washed off, leaving only the forward strands for sequencing.
The actual sequencing is also conducted in an iterative process. Starting from sequencing primers that are attached to the 3' ends of the templates, the reads are generated. To this end, fluorescence-labeled nucleotides (dNTPs) are added to the flow cell and compete for incorporation by DNA polymerases into the growing chains. After a matching dNTP is added to the backbone, the unbound nucleotides are rinsed, and the clusters are excited by a light source. A characteristic fluorescent signal is emitted from each cluster as each type of nucleotide is labeled with a specific fluorophore. Based on the emission wavelengths and signal intensities captured by an optical device, the incorporated nucleotides per cluster can be determined in a process called base calling. By this, millions of clusters can be sequenced simultaneously. The incorporated nucleotides contain reversible blocking groups to prevent the nascent chains from being extended in an uncontrolled manner. In a subsequent step, these terminators and the fluorophores are cleaved and the cycle starts over until the desired read length is achieved.
In a paired-end setting, i.e., in cases where a fragment is sequenced from both ends, the first read product is washed away and another round of sequencing is conducted. To this end, one more round of bridge hybridization and strand complementation is performed and the reverse strands are used for sequencing. The pairing of the reads provides additional spatial information that can be advantageous for further analysis steps (cf. **Section 3.1.2.3**). Separation distances can range from short inserts (200 to 500 bp) to long insert mate pairs (2 to 5 kb) [175].
Depending on the question at hand, the sequence reads are either assembled to a novel genome or aligned to a reference genome. Having paired-end reads helps to resolve ambiguous alignments and chromosomal rearrangements, such as insertions, deletions, and inversions [176]. Finally, various types of downstream analyses like variant calling, i.e., the identification of variations of the considered sample from the reference genome, can be performed [177]. Please refer to **Section 3.1.2.3** for an overview of the respective processing steps.

3.1.2.2 Other sequencing technologies and applications

High-throughput sequencing is a broadly used technique that can be employed to answer various types of research questions. In the following, several major types of applications will be outlined.

Targeted sequencing: Besides Whole Genome Sequencing (WGS), also targeted sequencing approaches like exome sequencing and panel sequencing are commonly applied. Exome sequencing, sometimes also called Whole Exome Sequencing (WES), focuses on the sequencing of the protein-coding regions in the genome (i.e., the exons), which only cover about 1% of the human genome [178]. Hence, for applications like the identification of genetic variants affecting protein sequences, WES is a much more time- and cost-efficient alternative to WGS [179]. Exome sequencing protocols are similar to those of WGS, with the additional step of exome capturing prior to amplification. Most commonly, biotinylated oligonucleotide probes that selectively target exons (so-called 'baits') are used for this purpose. After hybridization, the probe-DNA hybrids are seized by magnetic streptavidin beads and enriched via amplification [180]. An even more targeted approach to sequencing is Gene Panel Sequencing, in which only

up to several hundred genes are considered and which is typically used in clinical practice to probe for a predefined set of specific diagnostics- or treatment-relevant variants [181].
The drastically reduced number of bases that have to be sequenced in targeted sequencing approaches in comparison to WGS allows for an increased depth of sequencing at still relatively low costs. Sequencing coverage (i.e., the average number of unique reads that include a given nucleotide in the assembled sequence) thereby can be increased from low fold coverages of around 30x (as typically used in WGS) to high (typically at 100-200x) and ultra-high coverages (typically at 500-1000x) for exome and panel sequencing, respectively. By this, also low-frequency alterations can be reliably identified [182].

RNA sequencing: In addition to the elucidation of a sample's genomic features, high-throughput sequencing techniques can also be applied for transcriptome profiling via a class of approaches called RNA Sequencing (RNA-Seq) [183]. Similar to the targeted sequencing approaches described above, RNA-Seq requires an initial step of so-called 'target enrichment': Depending on the entities of interest, mRNA or different types of non-coding RNA need to be extracted from the cell lysate. To this end, magnetic beads with bound poly(T)-oligonucleotides can be used to capture only the mRNA from a cell lysate of interest by hybridizing to the poly(A)-tails of mature mRNAs. The mRNA molecules can then be sequenced using, in principle, any sequencing technology. In a next step, the obtained reads are either mapped to a reference genome or used for *de novo* transcriptome assembly. The coverage of the mapped reads within a gene can be used for quantification of the corresponding gene's expression (cf. **Section 3.1.2.3**). RNA-Seq overcomes some of the shortcomings of hybridization-based methods like the ones described in **Section 3.1.1**, as it is not limited to detecting transcripts that correspond to known genomic sequences, does not suffer from cross-hybridization bias, and provides a broader dynamic range of detection [184].

Bisulfite sequencing: Aside from the use of methylation microarrays, genome-wide methylation patterns can also be assessed via bisulfite sequencing [185]. Analogously to the sample preparation for methylation microarrays described above, the presence or absence of a methyl group at the fifth carbon position of cytosine pyrimidine rings is 'encoded' into the DNA via sodium bisulfite treatment. By this, unmethylated cytosines are deaminated to uracils, which are read as thymine when sequenced, while methylated cytosines remain unaltered and hence are read as cytosines. In comparison to hybridization-based and other methods, bisulfite sequencing entails the advantage of single-nucleotide resolution [186].

Chromatin immunoprecipitation sequencing: Another prominent application scenario for sequencing is Chromatin ImmunoPrecipitation Sequencing (ChIP-Seq), which is a combination of chromatin immunoprecipitation and deep sequencing to determine the occupancy of DNA with binding proteins [187]. Here, chromatin-binding factors and regulatory elements like transcription factors, histone modifications, or RNA polymerases are cross-linked to their bound DNA (e.g., via formaldehyde treatment). After cell lysis, the isolated chromatin is fragmented and the protein-DNA complexes are captured from the lysate via immunoprecipitation using specific antibodies. Sequencing of the DNA fragments, to which the corresponding regulators were bound to, allows for the identification of binding sites and motif discovery [188].

3.1.2.3 Processing of sequencing data

Similarly to the processing of microarray data (cf. **Section 3.1.1.3**), several image capturing and analysis steps are required for base calling, i.e. for the identification of the succession of nucleotides in the sequenced fragments. These initial processing steps are typically already performed by the sequencing machine [189]. The subsequent primary analysis of the obtained raw sequencing data is again a multistep process that, depending on the research question at hand, either provides a *de novo* genome (or transcriptome) assembly or uses a reference genome (or transcriptome) for further analysis. In the latter case, which we will focus on in the following, analysis workflows typically result in lists of somatic and/or germline mutations or scores of (differential) gene expression. The different analysis steps are conducted in elaborate pipelines that vary depending on the sequencing method used and the aim of the analysis. The main analysis steps, which will be outlined below, are quality control, read mapping, and either variant calling/annotation or copy number quantification for genome sequencing scenarios or RNA quantification in the case of RNA sequencing. There are numerous tools for each of these analysis steps [190–192].

Quality control: Typically, the sequencing machine yields information about the identified reads, i.e. the sequenced fragments of DNA or RNA, in the form of `FASTQ` files (see **Section A.1.2** for details on the file format). Besides the actual sequencing information, `FASTQ` files contain scores indicating the quality of each base call as estimated by the sequencing machine. These quality scores, called Phred scores Q_{ij}, indicate the probability of the reported nucleotide n_{ij} to be a sequencing error and are defined as follows:

$$Q_{ij} = -10 \log_{10} \left(p_{ij} \right),$$

with p_{ij} being the probability of an incorrect base call of the j-th base in read i. Typically, base calling quality decreases towards the end of a read [193]. Hence, the Phred scores can be used to trim reads such that only high-confidence nucleotides are considered for further analysis. Another class of artifacts potentially contained in the raw data is introduced by sequencing adapters. Sequencing adapters are short nucleotide sequences ligated to the genomic fragments for amplification and sequencing as part of the experimental protocol (cf. **Section 3.1.2.1**). The sequences of the used adapters or fragments thereof are sometimes erroneously contained in the obtained read sequences. Hence, in order to remove low-quality bases and adapter artifacts from the raw sequencing data, tools like Trimmomatic [194] or Flexbar [195] should be included in corresponding sequencing pipelines.

Read mapping: The experimental principle of 'shotgun sequencing' as employed by most sequencing techniques provides reads that do not contain any information about where in the genome they are originating from. Hence, the reads have to be mapped to a corresponding reference genome, which is provided for numerous organisms by several databases (cf. **Section 3.2.1**). The task of read mapping corresponds to the solving of an approximate string matching problem, i.e. the search for occurrences of a read within a reference sequence while allowing for some mismatches and gaps between the two. This tolerance is required as (i) the reads contain sequencing errors that have to be accounted for and (ii) the

genomic sequence of a sample under investigation is expected to differ from the reference genome [196]. This difference becomes especially evident when mapping reads originating from the genomes of cancerous cells, as these are likely to contain numerous germline as well as somatic aberrations.

There are a plethora of tools for read mapping [197], of which the Burrows-Wheeler Aligner (BWA) [198] and Bowtie2 [199] are two of the most popular ones. In order to overcome the combinatorial explosion of putative alignments, the majority of these tools apply indexing and filtering techniques to quickly scan and reduce the search space [200]. In this context, indexing means the preprocessing (e.g., via Burrows-Wheeler transformation [198]) and representation of the reference genome, the reads, or both in specific data structures (e.g., suffix array [201], FM-index [202]) that trade a much faster identification of putative alignment positions off against a larger memory consumption. Filtering strategies rely on the identification of short regions in the reference genome (k-mers) that perfectly match to a small fraction of the read (the 'seed'). Only for those regions, the whole read is being aligned using optimized local dynamic programming-based algorithms [203, 204], which also provide a measure of alignment quality.

Besides the intrinsic difficulty of the approximate string matching problem, additional challenges arise when mapping reads obtained from RNA sequencing experiments. When aligning reads obtained from transcripts, the alignment potentially has to be split along intron-exon boundaries. A commonly used read mapping tool for the identification of such spliced alignments is TopHat2 [205].

Although bioinformatics approaches can tackle various challenges arising in the context of read mapping, the inherent structure of the DNA hinders unambiguous mappings in many cases. The genomes of a broad range of species, from bacteria to humans, contain a high number of repetitive regions ('repeats') [206]. Reads originating from these repetitive regions cannot be unambiguously mapped by the read aligner. However, in the case of paired-end sequencing (cf. **Section 3.1.2.1**), the alignment tool can utilize the fact that the paired sequences were derived from the two ends of a fragment of (roughly) known length.

A current standard for the output of read mapping tools are the Sequence Alignment/Map format (`SAM`) or its binary version, the Binary Alignment/Map format (`BAM`). For details on these file formats, please refer to **Section A.1.3**.

After the actual alignment step, some pipelines include an optional step of re-aligning reads around indels. This step is based on the observation that reads harboring indels are especially prone to suboptimal alignments. To reduce this bias, a re-alignment around these positions can be performed, for example using the Genome Analysis ToolKit (GATK) [207, 208].

Variant calling: In its most general form, variant calling is the process of identifying sequence differences between the mapped reads of a sample under investigation and the reference genome. These differences can serve as the basis of, for example, an examination of the molecular causes of genetic disorders or the identification of contributors to the initiation and progression of complex diseases like cancer.

In an ideal world, variant calling and the related process of genotype calling, i.e. the determination of a variant's zygosity to be either heterozygous (only one allele affected) or homozygous (both alleles affected), could be performed by simply assessing the frequency of the variation across all reads mapped to the considered genomic location. Yet, preprocessed

sequencing data can exhibit high error rates that are due to various factors, including base calling inaccuracies and ambiguous alignments. As a consequence, frequency-based approaches using fixed thresholds are only feasible for deep sequencing data, i.e. for samples in which specific genomic locations are, on average, covered by a sufficiently large number of reads [209]. As a remedy, there are various tools and methods based on heuristics or probabilistic models that try to reduce and to quantify the uncertainty associated with variant and genotype calling by considering prior information such as Phred quality scores, read coverage, and allele frequencies. Prominent examples of such tools are MuTect [210], SomaticSniper [211], and VarScan2 [212]. Solving the task of variant calling has the potential to yield a comprehensive overview of various types of local genomic aberrations, including base substitutions and short insertions or deletions, as well as copy number variations or even larger structural aberrations [213, 214].

However, as already discussed in **Section 2.2.1.1**, not all deviations from the reference genome are specific to diseased cells: with respect to the reference genome, samples can contain germline mutations as well as somatic mutations. In the context of oncology, variant calling oftentimes mainly refers to the identification of somatic mutations in the cancer genome, as (some of) these variants are assumed to drive the progression of the tumor. A differentiation of mutations in somatic and germline variants can be achieved in three ways: either via (i) the comparison to a healthy control sample from the same patient, (ii) the consideration of allele frequencies, or by (iii) filtering for known somatic variants using variant databases. The first, and preferable, approach requires the availability of a healthy control sample from the same individual to distinguish somatic from germline variations. The two samples then can either undergo variant calling independently and the mutations in the intersection of both are considered as germline mutations or they are jointly analyzed [215]. Another approach for the case that no patient-matched normal control is available is the consideration of allele frequencies. In combination with information about tumor purity, i.e. the estimated fraction of tumor tissue in the sample, they can be used to predict a variant's zygosity and whether it is assumed to be a somatic or germline mutation [216, 217]. Lastly, there are numerous databases that contain knowledge about a plethora of variations and whether they have previously been reported as germline variants or being cancer-associated (cf. **Section 3.2.3**). While being independent from a control tissue sample, this approach has the major drawback that it is dependent on the quality and completeness of the used databases and that low-frequency variants will hardly be distinguishable from sequencing noise.

While somatic mutations in cancer cells are typically the primary focus for the identification of suitable targeted treatment options, the consideration of germline aberrations should not be neglected, as they can play essential roles in the identification of predispositions for certain diseases or the assessment of drug-processing enzymes (cf. **Section 2.3**).

The results of variant calling pipelines are commonly represented in the Variant Call Format (`VCF`) (see **Section A.1.4** for additional details on the file format), which serves as input for the subsequent variant annotation step.

Variant annotation: The sequencing of cancer genomes can reveal a large number of somatic (and/or germline) mutations, all of which have a different degree of impact (if any) on the disease phenotype (cf. **Section 2.2.1.1**). The process of variant annotation aims at aggregating

and reporting relevant information to a given genomic alteration and is an essential step in sequencing analysis pipelines to facilitate the interpretation of the detected mutations [191].
A first step in pinpointing those mutations that are relevant for disease initiation and progression is the identification of the genomic region or gene affected by a given variation and its functional effect. There are several general-purpose tools like ANNOVAR [218], SnpEff [219], or Ensembl's Variant Effect Predictor (VEP) [220] that categorize variants based on their predicted impact on protein function. For example, VEP distinguishes more than 30 so-called 'consequence types', i.e. different effects a mutation can have on the genomic location or protein it occurs in (cf. **Section A.2**). While severe mutations like a frameshift or nonsense mutations are generally assumed to destroy the structure and hence functionality of the affected protein, the functional impact of other types of consequences is non-obvious. To this end, the annotations made by the aforementioned tools can be extended by additional classifications of the consequences in terms of their predicted severities. For non-synonymous variants, tools like Sorting Intolerant From Tolerant (SIFT) [221] or Polymorphism Phenotyping v2 (PolyPhen-2) [222] can be used. SIFT is a tool that predicts the effect of a mutation on the affected protein's function based on sequence homology and the physicochemical characteristics of the exchanged amino acids. PolyPhen-2 uses a variety of sequence-based and structure-based features to predict the probability that an amino acid substitution has a damaging effect.
Although these annotations can provide first insights into the potential role of specific mutations in the disease context, they were developed for general applications and hence lack cancer-specific annotations that could aid downstream analysis and interpretation. To this end, various databases containing cancer-specific annotations like ClinVar [223], CIViC [224], COSMIC [225], or dbSNP [226] can be employed. Additional details on these databases can be found in **Section 3.2.3**.

Copy number quantification: High-throughput sequencing data can also be used to assess copy numbers within a sample. In contrast to aCGH arrays, which were presented in **Section 3.1.1.2** and which identify relative copy number changes, there are efforts to derive absolute copy numbers from whole-genome or whole-exome sequencing data. There are numerous methods that, to this end, typically consider read counts or investigate distances between read pairs. Examples for such tools are CNVnator [227], CNV-seq [228], or VariationHunter [229].

RNA quantification: As an alternative to microarray experiments for the quantification of transcript levels (cf. **Section 3.1.1**), gene expression can also be measured using RNA sequencing techniques (cf. **Section 3.1.2.2**). As already indicated in the previous paragraphs, the quality control and read mapping steps are generally similar to those of DNA sequencing data. However, in order to account for splicing events that may prevent RNA reads from being consecutively aligned to the reference genome, specialized tools can be employed. Popular RNA aligners are TopHat2 [205] and STAR [230]. After the (spliced) alignment to the transcriptome, the coverage of RNA reads across a gene can be used to infer its expression level. To this end, the read counts are normalized with respect to various confounding factors, including total read count and gene length [231]. Here, commonly used tools include HTSeq [232] and featureCounts [233]. Expression levels are quantified by counting the number of reads that have been mapped to a

locus of interest. In order to make RNA-Seq results comparable between runs, absolute counts are typically converted to Fragments Per Kilobase of Exon per Million Reads Mapped (FPKM) [234] or related measures [235]:

$$FPKM_i = \frac{q_i}{\frac{l_i}{10^3} \cdot \frac{\sum_j q_j}{10^6}} = \frac{q_i}{l_i \cdot \sum_j q_j} \cdot 10^9,$$

where q_i is the number of fragments mapped to a specific gene i, the length of which is denoted as l_i and the total number of mapped fragments over all genes j is given by $\sum_j q_j$. In contrast to data derived from microarray experiments, which are typically assumed to yield normally distributed values, tools for estimation of differential expression based on RNA-Seq count data (e.g., Cufflinks [236], DESeq [237]) typically assume an underlying Poisson or negative binomial distribution.

3.1.3 Mass spectrometry

The last experimental technique described in this chapter - and an emerging means of protein and metabolite quantification - is mass spectrometry. Similar to microarray and sequencing technologies, mass spectrometry is based on the identification and quantification of the molecular fragments making up the biological entities under consideration, which are in the case of mass spectrometry proteins or metabolites. To this end, mass-to-charge ratios and intensities of the respective ionized molecular fragments are measured and analyzed. There are various protocols and types of mass spectrometry, of which the commonly used technique of Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS) for proteins is described in the following.

3.1.3.1 Experimental principle

LC-MS/MS is based on the stepwise selection, fragmentation, and mass analysis of proteins and peptides of interest [238]. **Figure 3.3** gives an overview of the processing steps.

In a first step, proteins are extracted from the cells of interest. Depending on the type of analysis to be conducted, the mixture of proteins in the lysate can be separated by molecular weights and/or isoelectric points (e.g., via 1D or 2D gel electrophoresis). The subsequent analysis of single bands reduces the complexity and increases the accuracy of the mass spectrometry. The selected proteins are proteolytically digested (e.g., using trypsin) to obtain a characteristic mixture of peptides due to the used enzyme's specific cleavage sites [239].

In order to improve both, the sensitivity and specificity of the MS, the mixture of peptides is separated in a process called Liquid Chromatography (LC). In LC, the sample is pumped through a stationary phase (e.g., a silica gel) using an organic solvent as the mobile phase. By this, the mixture's constituents are separated by polarity due to their relative affinities to the mobile and the stationary phase, respectively [240]. The peptides eluting from the LC are subsequently ionized. To this end, soft-ionization techniques like electrospray ionization can be used, which generate ions from macromolecules without breaking their chemical bonds [241].

The charged peptides undergo the first stage of mass analysis, which separates them based on their mass-to-charge ratio. This first run results in a mass spectrum from which typically peptides of a particular mass-to-charge ratio are selected for further analysis.

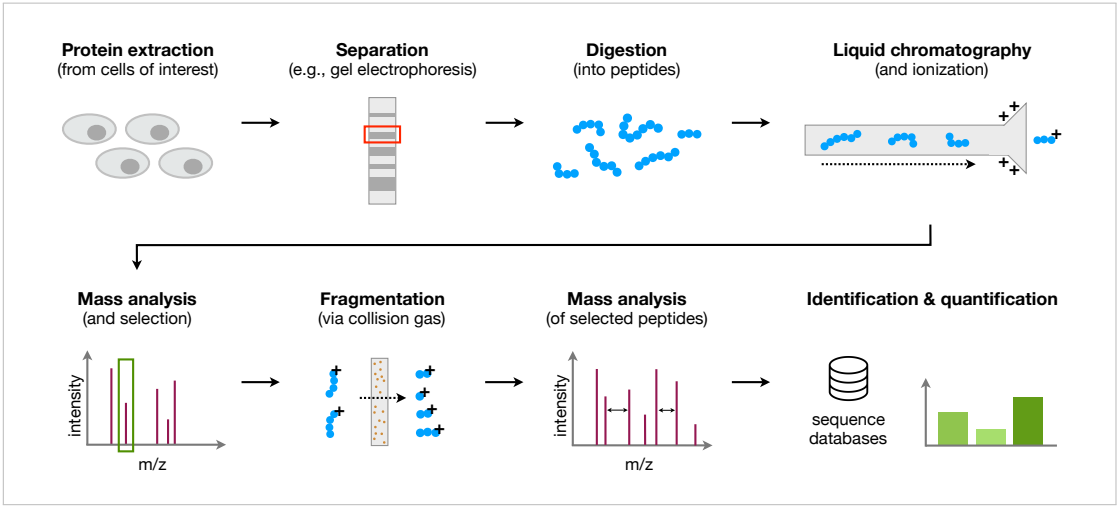


Figure 3.3 **Overview of LC-MS/MS workflow**. After extraction and separation by gel electrophoresis, a subset of proteins are enzymatically digested and the mixture is furthermore separated by liquid chromatography. As soon as the peptides elute from the chromatography column, they are ionized and undergo a first mass analysis (MS). Peptides at a particular mass-to-charge ratio (m/z) are selected for further fragmentation and a second run of mass analysis (MS/MS). Based on the observed spectrum, the originating proteins and their abundances can be inferred. The database icon was obtained from [15].

After further fragmentation of the 'precursor ions' (e.g., via collision-induced dissociation [242]) into 'product ions', the ions undergo a second round of mass analysis. The resulting characteristic pattern of peaks and, in particular, their relative differences can be used to identify the amino acid sequence of the analyzed peptide. To this end, the observed mass spectrum is queried against a database of theoretical masses of *in silico*-digested peptides (based on known enzyme specificities). After the identification of the different individual peptide sequences, these sequences are used to infer which protein(s) were present in the sample [243]. Using mass spectrometry, not only different proteins can be identified and distinguished, but also the presence or absence of post-translational modifications (PTMs) (cf. **Section 2.1**), as they yield observable peak distances in the MS/MS run. This allows for the elucidation of signal transduction processes based on the assessment of protein phosphorylation states, but also other types of PTMs like ubiquitination, sumoylation, or glycosylation can be assessed [244].

3.1.3.2 Quantitative proteomics

Besides the identification of proteins, also their abundance can be approximated from the detected signal intensities, i.e. the height of the corresponding peaks in the mass spectrum [245]. However, the observed peak heights are affected by a variety of perturbing factors, e.g. varying efficiencies in protein digestion or alterations in the degree of ionization of a peptide due to coeluting substances ('matrix effects'). These perturbing factors can strongly vary between experimental runs and hence hinder a relative quantification of proteins across experiments [246]. As a remedy, internal references can be used for quantification, for example based on stable isotope labeling. Corresponding methods relate the protein expression of two samples to each other, one of which is isotopically labeled. This approach relies on the theory that a stable isotope-labeled peptide possesses the same chemical properties as its native counterpart and hence is also affected to the same extent by the bias-inducing factors mentioned above [96]. Yet,

the labeled and unlabeled form of a peptide display a specific difference in their mass-to-charge ratios in the resulting mass spectrum, which facilitates accurate quantification via the relation of their relative signal intensities.

There are several methods for isotope quantification, which mainly differ by the way in which stable isotope labels are introduced into the proteins or peptides of interest. Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC) [247] is one of the most prominent methods for isotope quantification and follows the approach of metabolic labeling.

For SILAC, two populations of cells are grown in cell culture, one of which is cultivated in an isotope-labeled medium, in which lysine and arginine are substituted by their heavy isotope counterparts. Typical isotopes used in SILAC to replace the respective atoms in light amino acids are ^{2}H, ^{13}C, ^{15}N, and ^{18}O [248]. The substitution within the cells occurs via protein turnover and due to the fact that the cells have to ingest essential amino acids (i.e., amino acids that cannot be synthesized by the cells themselves) from the medium. Although arginine actually is a non-essential amino acid in adult vertebrates, studies have shown that arginine becomes essential in cell cultures [249, 250]. In combination with the fact that trypsin specifically cleaves after lysine and arginine, this ensures that nearly every peptide originating from the heavy sample will be isotopically labeled when analyzed using mass spectrometry. SILAC can be used to investigate the effects of perturbations on protein expression. To this end, one of the populations is perturbed, e.g. via the treatment with a drug. The cells from the two populations are then lysed and mixed in a 1:1 ratio, followed by the combined processing of the sample and MS analysis.

For comparisons across several samples and conditions, a modification of the classic SILAC-approach, called spike-in SILAC [251] can be used. Here, cells grown in heavy media are used as a reference in each sample, yielding normalized abundance ratios for every sample that cancel out when relating two samples to each other.

3.1.3.3 Processing of mass spectrometry data

The processing of mass spectrometry data to yield qualitative and quantitative information about the contained proteins (or metabolites) is a multistep procedure, which will be briefly described in the following paragraphs for the proteomics case. For each step within these pipelines, there exist numerous tools that are optimized for different types of analyzes [252–254]. Many of these tools are also part of comprehensive frameworks for the processing of mass spectrometry data like OpenMS [255], Mzmine2 [256], or MAVEN [257].

Raw data filtering: Fragmentation-based approaches like tandem mass spectrometry yield a large variety of fragment ions, whose corresponding peaks in the mass spectrum are commonly impaired by spectral noise. In order to increase the accuracy of peptide identification, an initial noise filtering step is performed to separate the actual signal from background noise (e.g., induced from chemical or instrumental interference) [258]. To this end, the peaks with the weakest intensities (i.e., background noise) are filtered out.

Peak detection and identification: Before peptides are identified from MS/MS spectra, typically a (computational) deisotoping step is performed. Due to the fact that an analyte can sometimes contain naturally occurring rare isotopes (e.g., ^{13}C), there might be additional minor peaks besides the major monoisotopic peak (that occurs at the theoretical mass) in the mass

spectrum. Deisotoping algorithms like the ones provided by Decon2LS [259] or PeakSelect [260] try to group the isotopic peaks with the corresponding monoisotopic peak.
The next step is to identify the peptides that correspond to the (monoisotopic) peaks in the filtered spectrum. To this end, one or several databases are used to compare the experimental features against patterns of theoretically digested proteins using tools like SEQUEST [261] or Mascot [262]. After the identification of the contained peptides, they have to be mapped to their originating proteins. In contrast to simple organisms, where most peptides can be uniquely mapped to one protein, the problem becomes much more complex for higher eukaryotes [263] as many peptides could stem from several proteins, thus leading to ambiguous protein assignments. A well-known tool to perform this inference is ProteinProphet [264], which tries to identify the minimal set of proteins that can explain all the observed peptides.

Quantification: Besides the identification of proteins contained in a sample, for many research scenarios it is of high interest to also measure and compare protein abundances to identify differences between two samples or sample groups. In label-free cases where each sample is analyzed individually, matching peaks across experiments have to be identified. The heights of the corresponding peaks are used to determine the relative peptide abundances between the samples. However, in general, this approach is less accurate than methods employing stable isotope labeling (cf. **Section 3.1.3.2**). In these labeled approaches, the relative quantity of a specific protein can be determined from their relative signal intensities using tools like ASAPRatio [265] or MaxQuant [266].

3.2 Reference databases and resources

High-throughput technologies as the ones described in **Section 3.1** produce large amounts of complex, noisy, and heterogeneous data that require the use of bioinformatics methods in combination with biological domain knowledge to obtain novel insights into disease biology and putative treatment options. In this context, databases and other resources act as an enabling factor for feature annotation and systems biology integrative analyses. There is a large and steadily increasing number of databases that capture various types of biological knowledge, facilitating a multi-faceted investigation of data sets of interest [267].
This plethora of databases can be classified in various ways: for instance, some databases contain general-purpose information, while others are specific to certain organisms or diseases. Also, they can be distinguished in terms of accessibility, cost of use, and whether or not they are continuously updated. Another essential difference between and sometimes even within databases is the level of evidence provided for the entries: some databases only provide validated information, while others also contain derived or predicted entries.
In the following, we will present three main classes of databases and resources especially relevant to basic cancer research and translational medicine. To this end, we here focus on the primary types of information provided by the respective databases, acknowledging that a sharp partitioning is hardly achievable due to the complexity and comprehensiveness of current databases.

3.2.1 Entity-related databases

This class of databases entails various types of resources that provide biological context to specific biological entities like genomic locations, genes, or proteins. Similar to the 'flow of genetic information' (cf. **Section 2.1**), they range from reference genomes and catalogs of genes over information on regulatory elements to protein sequence and structure databases.

The National Center for Biotechnology Information (NCBI) provides reference genomes for a multitude of organisms in their Reference Sequence (RefSeq) database [268]. The Gencode database, on the other hand, focuses on the identification and classification of genes in *Homo sapiens* and *Mus musculus* and provides corresponding reference genomes for those two organisms [269].

One of the most comprehensive databases for gene-specific information is NCBI's (Entrez) Gene database [270]. The provided database records include general-purpose information like alternative gene names, summary descriptions, genomic localizations, tissue-specific expression patterns, as well as links to literature and other external resources. GeneCards is another gene-centric resource that integrates various types of information from a large variety of external databases [271].

The HUGO Gene Nomenclature Committee (HGNC) [272] provides approved human gene names and short-form abbreviations ('gene symbols') for almost 40,000 human loci, including protein-coding and non-coding genes, as well as corresponding identifiers in other nomenclatures that can be used for identifier mapping.

Going beyond gene-level annotation, the Ensembl project [273] covers various types of entities, ranging from single exons over transcripts to genes and their respective gene products. Each entity is provided with a stable alphanumeric identifier and includes information about genomic locations for various reference genomes and cross-references to related entities.

Besides Ensembl, the Universal Protein Resource (UniProt) is a major resource for protein annotation, including details on the three-dimensional structure, post-translational modifications, and functional annotations [274].

For the elucidation of regulatory mechanisms concerning gene expression, there is a large variety of databases that provide information on regulatory elements like transcription factors and their target sites (e.g., TRANSFAC [275], JASPAR [276], or ENCODE [277]) and miRNAs (e.g., miRBase [278], miRCarta [279], or miRWalk [280]).

3.2.2 Pathway-related databases

Nowadays, there are various resources and databases that try to structure the complex molecular circuitry within cells by providing functional annotations to gene sets, biological processes, and pathways.

The most comprehensive and renowned resource for the functional annotation of gene sets is the Gene Ontology (GO) database [281]. GO provides a hierarchy of controlled vocabulary terms to describe the roles of genes and their respective gene products concerning their molecular function(s), the biological process(es) they are involved in, and the cellular component(s) in which molecular events occur. While GO is a valuable resource for gene set annotations, it does not provide information about the specific interactions (i.e., chemical reactions and binding events) between genes, proteins, non-coding RNAs, and metabolites that together give rise to a

cellular function. The Kyoto Encyclopedia of Genes and Genomes (KEGG) comprises several databases that provide details on regulatory signaling cascades, biochemical reactions, and additional information about the involved genes and proteins [282]. The encompassed KEGG PATHWAY database provides a collection of manually created pathway maps that represent knowledge of molecular interaction networks, including additional knowledge about interaction types like activation, inhibition, phosphorylation, and dephosphorylation.
Similarly to KEGG, Reactome is an expert-curated and peer-reviewed pathway database [283]. Reactome's data model covers protein, protein complexes, metabolites, and their respective interactions. WikiPathways [284] is another resource providing interactive biological pathway maps via an open, collaborative platform.

3.2.3 Disease- and drug-related databases

For personalized medicine and decision support, the incorporation of disease-specific, as well as pharmacologic and pharmacogenomic knowledge is of great importance.
The Single Nucleotide Polymorphism Database (dbSNP) is a comprehensive resource for annotations of genetic variations, including population-specific frequencies, links to related publications, and information about the pathogenicity of mutations, i.e., their predicted role in human disease [226]. Unlike indicated by the name, dbSNP does not only contain Single Nucleotide Polymorphism (SNPs), i.e., single nucleotide substitutions that occur in at least 1% of a population, dbSNP also covers other types of short sequence variation (e.g., small insertions or deletions) that occur frequently enough to be termed polymorphic [285]. Similarly to dbSNP, the ClinVar database [223] describes the relationships between human genetic variations and disease phenotypes. Moreover, the clinical significance of aberrations and supporting evidence is provided whenever available.
The Catalogue Of Somatic Mutations In Cancer (COSMIC) focuses on the potential role of mutations in human cancer. COSMIC is currently the most comprehensive publicly available resource for expert-curated somatic information and their relation to human cancers [225].
Another resource for the evaluation of the clinical relevance of inherited and somatic variants in cancer is CIViC, the Clinical Interpretation of Variants in Cancer database [224]. CIViC is an open access, expert-crowdsourced knowledgebase that provides several levels of evidence for the assessment of the predictive power of aberrations with respect to the treatment with various drugs.
As already discussed in **Chapter 2**, tumors can potentially contain various types of (epi-)genomic variations in numerous genes, not all equally contributing to cancer initiation and progression. Databases like the Integrative Onco-Genomics database (IntOGen) [286] and DriverDB [287] can help to prioritize genes for further investigation by providing information about known and putative driver genes and corresponding aberrations for a variety of cancer types.
Information about a broad range of drugs and their molecular targets is provided by various resources like DrugBank [288], the Drug Gene Interaction Database (DGIdb) [289], and the Therapeutic Target Database (TTD) [290]. DrugBank is a comprehensive resource that combines detailed drug data (i.e., chemical, pharmacological, and pharmaceutical properties of a drug) with additional details on drug targets and relevant ADME genes (i.e., drug-processing

enzymes, transporters, carriers). The latest release of DrugBank (Version 5.1.1) contains almost 12,000 drug entries, covering all stages of development from experimental to approved. DGIdb provides a comprehensive collection of drug-target interactions gathered and consolidated from the literature and a broad array of databases and web resources. In addition to drug-target interactions, TTD additionally provides information about target-affiliated biological pathways. The data provided by these databases can be employed to find existing drugs for a specific target (e.g., to assess putative points of intervention within a pathway or gene set) or to identify all known targets of a drug under consideration.

(Epi-)genomic and transcriptomic aberrations in molecular drug targets can significantly impact the efficacy and toxicity of drugs. The consideration of predictive biomarkers (CDx) and pharmacogenomic interactions (PGx) is hence an integral part of personalized medicine. There exist a variety of databases that focus on CDx and PGx relationships and which vary in their comprehensiveness and provided levels of evidence.

The FDA, for instance, only provides a list of approved and ready-to-use companion diagnostics that test for specific aberrations determining treatment eligibility [291].

A broader set of annotations is provided by OncoKB [292], which covers more than 4,000 genomic alterations for several dozens of drugs. These alterations are classified into four levels of clinical evidence.

An even more comprehensive resource providing putative pharmacogenomic effects is the Genomics of Drug Sensitivity in Cancer (GDSC1000) database [293]. GDSC1000 contains the predicted effects of aberrations on drug efficacy based on drug sensitivity measurements for a wide range of drugs across numerous types of cell lines.

Besides databases as the ones described above, there are also numerous resources that provide disease-specific multi-omics data sets and, in some cases, also drug sensitivity information that can be used to infer cancer type-specific patterns of (epi-)genetic and molecular aberrations and determinants of drug sensitivity. The Gene Expression Omnibus (GEO) [294], ArrayExpress [295], NCBI's Sequence Read Archive (SRA) [296], and the European Genome-Phenome Archive [297] are examples of comprehensive resources that provide various types of omics data across diseases and sample conditions. The currently largest multi-omics data set on primary and metastatic tumor samples is provided by The Cancer Genome Atlas (TCGA) [298]. Other resources like GDSC1000 [293] and the DREAM7 data set [299] combine multi-omics measurements for cell lines of various types with drug sensitivity measurements for large panels of drugs.

3.3 Detecting deregulated genes and processes

The preprocessed data obtained from the high-throughput experimental technologies described in **Section 3.1** is of high dimensionality and complex nature. Thus, in order to extract meaningful biological information from these data, appropriate statistical and computational means need to be applied. In this section, we will briefly introduce the statistical concept of hypothesis testing and significance (**Section 3.3.1**), followed by an overview of various methods for the detection of deregulated genes (**Section 3.3.2**) and biological processes (**Section 3.3.3**).

3.3.1 Hypothesis testing and significance

Hypothesis testing is a commonly used technique of inferential statistics for the assessment of properties of a population based on an observed sample [300, 301]. The following definitions and explanations are based on [302].
A property of such a population is typically represented as an unknown parameter $\theta \in \Omega$ of the underlying probability distribution, i.e. the probability distribution from which the observed sample was drawn. The parameter space Ω can be partitioned in two disjoint subsets Ω_0 and Ω_1 ($\Omega = \Omega_0 \cup \Omega_1$) that give rise to two hypotheses about the value of θ: the so-called 'null hypothesis' (H_0) and the 'alternative hypothesis' (H_1):

$$H_0 : \theta \in \Omega_0$$
$$H_1 : \theta \in \Omega_1$$

While H_1 typically corresponds to the research hypothesis of interest (see also the example below), H_0 represents the complementary hypothesis. As we cannot directly assess the correctness of H_1, we instead consider the probability of obtaining the observed sample under the assumption that H_0 is true. If this probability is sufficiently small, we can reject the null hypothesis H_0 in favor of H_1. In order to determine whether or not to reject H_0, a test procedure δ_c is used. The test procedure (decision rule) δ_c is a function that determines whether or not to reject H_0 for a given random sample $x = (x_1, x_2, ..., x_n)$. To this end, a so-called 'test statistic' is employed. A test statistic $T : \mathbb{R}^n \to \mathbb{R}$ is a real-valued function that maps the made observations to a scalar value. If the result of the test statistic $T(x)$ exceeds the critical value c defined by the test procedure δ_c, H_0 will be rejected. By this, the test procedure δ_c partitions the sample space S (i.e., the set of all possible random samples x) into two subsets: the so-called 'critical region' (CR), for which the test procedure δ_c will reject H_0 and the complementary subset (CR^c) for which H_0 will not be rejected.
The probability that the test procedure δ_c will reject H_0 is defined by the so-called 'power function' $\pi(\theta|\delta_c)$:

$$\pi(\theta|\delta_c) = P(x \in CR|\theta) \quad \forall\theta \in \Omega$$

In hypothesis testing, two classes of errors can occur: type-I and type-II errors. A type-I error refers to the erroneous rejection of a true null hypothesis. The probability of making a type-I error is denoted by:

$$P\{\text{type-I error}\} = \sup_{\theta\in\Omega_0} (\pi(\theta|\delta_c))$$

Conversely, in cases where H_1 holds but H_0 is erroneously not rejected, a type-II error occurs. The probability of making a type-II error is denoted by:

$$P\{\text{type-II error}\} = \sup_{\theta\in\Omega_1} (1 - \pi(\theta|\delta_c))$$

The significance level α of a test procedure δ_c defines an upper bound for the probability of drawing the wrong conclusion by rejecting a correct null hypothesis, i.e. the probability of

conducting a type-I error:

$$\sup_{\theta \in \Omega_0} (\pi(\theta|\delta_c)) \leq \alpha$$

Typically, α is set to 0.01 or 0.05.
Finally, in order to determine whether or not to reject H_0 on the basis of a random sample x, suppose that the value of the test statistic is given by $t = T(x)$. For the test procedure δ_t that rejects H_0 if $T(x) \geq t$, we define the p-value p of this observation as:

$$\begin{aligned} p &:= \sup_{\theta \in \Omega_0} (\pi(\theta|\delta_t)) \\ &= \sup_{\theta \in \Omega_0} (P(T \geq t|\theta)) \end{aligned}$$

If the obtained p-value is smaller than the predefined significance level α, H_0 will be rejected in favor of H_1.

For illustrative purposes, consider the following example of replicated measurements of a gene's expression in two conditions X and Y, e.g. in diseased and in healthy state: The measurements yield two series of values $x = (x_1, ...x_n) \in \mathbb{R}^n$ and $y = (y_1, ..., y_m) \in \mathbb{R}^m$. For the assessment of whether there is a significant difference in the mean gene expression between the two conditions X and Y (i.e., whether or not the gene is differentially expressed), the null hypothesis would state that the two underlying populations have identical mean values ($H_0 : \mu_X = \mu_Y$), whereas the alternative hypothesis would state that there is a difference ($H_1 : \mu_X \neq \mu_Y$).
For the evaluation of these statements, a suitable test statistic T needs to be applied. For the given example, a possible test statistic would be the t-statistic (see also **Section 3.3.2**):

$$T(x_1, ..., x_n, y_1, ..., y_m) = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}} \sqrt{\frac{1}{n} + \frac{1}{m}}},$$

with $\bar{x}$ and $\bar{y}$ being the respective sample means and s_x^2 and s_y^2 the respective sample variances. The value of the test statistic T on the observed data is given by $t = T(x_1, ..., x_n, y_1, ..., y_m)$. The extremity of t under the null hypothesis can be used to judge the significance of the observation:

$$p = P(|T| \geq |t| \mid H_0)$$

This definition corresponds to a two-tailed testing scenario that is applicable in cases in which the alternative hypothesis embodies deviations from a reference in either direction. In contrast, statistical testing that examines only deviations in one of the two possible directions is called one-tailed or, more specifically, left-tailed ($p_l = P(T \leq t|H_0)$) and right-tailed ($p_r = P(T \geq t|H_0)$). Observations with sufficiently small p-values ($\leq \alpha$) indicate the observation is unlikely to occur under the null hypothesis. Hence, H_0 should be rejected in favor of H_1.

While p-values are the *de facto* standard when assessing statistical hypotheses, it has to be acknowledged that there has been an ongoing controversy around their informative value as they do not provide information about the actual biological relevance of the detected effect

[303, 304]. Hence, in addition to a p-value, the actual effect size should always be stated and considered when assessing the biological significance of a result. The effect size is the magnitude of the observed differences between groups. It can be measured, for example, as the difference between the means of the expression scores of a sample group and a reference group [305].

3.3.1.1 Multiple hypothesis testing

As established above, the significance level α describes the expected probability of making a type-I error when performing a single hypothesis test. However, in most studies and especially in the analysis of biological high-throughput experiments, a multitude of hypotheses are tested, e.g. for the assessment of differential expression of thousands of genes. Hence, when conducting $n = 1,000$ hypothesis tests at a significance level of $\alpha = 0.05$, we have to expect to obtain $\alpha \cdot n = 0.05 \cdot 1,000 = 50$ false-positive results (i.e., false discoveries) just by chance. Consider the case that a hypothesis test results in the rejection of H_0. The probability that this was the correct decision is $1 - \alpha$. Accordingly, the probability of making the correct decision in n hypothesis tests when always rejecting H_0 is given by $(1 - \alpha)^n$. Based on this observation, the Family-Wise Error Rate (FWER) is defined as the probability of making at least one false discovery in n different rejected tests:

$$p_{\text{fwer}} = 1 - (1 - \alpha)^n.$$

With an increasing number of tests n, this error rate increases and hence has to be accounted for. There are various methods to control the FWER, i.e. to adjust the single test type-I error in such a way that the overall type-I error rate remains below a given threshold. One of the most conservative methods is the Bonferroni correction [306]. Consider the case that n hypothesis tests were conducted, yielding n p-values $p_1, ..., p_n$. Instead of accepting p_i if it is smaller the significance threshold α, the corresponding observation i is only considered significant if $p_i \cdot n \leq \alpha$ holds. There are also other methods for FWER correction available that are slightly less strict while maintaining firm control of the FWER, e.g. as proposed by Šidák [307], Holm [308], and Finner [309].

An alternative to controlling the FWER is the consideration of the False Discovery Rate (FDR). The FDR is the expected proportion of false positives (type-I errors) among the rejected null hypotheses. Commonly used methods to control the FDR were presented by Benjamini and Hochberg [310] and by Benjamini and Yekutieli [311].

The Benjamini-Hochberg method adjusts each original p-value p_i. The adjusted p-values are typically called q-values q_i. Consider a list of n independent p-values sorted in ascending order $p_1 \leq p_2 \leq ... \leq p_n$. The q-value q_i for a p-value p_i is then computed as follows:

$$q_i = \begin{cases} p_i & \text{, if } i = n \\ \min\left\{q_{i+1}, \frac{n}{i} \cdot p_i\right\} & \text{, } \forall i \in \{n-1, ..., 1\} \end{cases}$$

Rejecting H_0 only for those observations i with $q_i \leq \alpha$, controls the FDR to be at most α (under the assumption of independent hypotheses being tested). An extension of this approach is the Benjamini-Yekutieli method [311], which controls the FDR under arbitrary dependence assumptions.

3.3.2 Detecting deregulated genes

The identification of differences between two states of interest (e.g., comparing diseased vs. control or treated vs. untreated) is a typical goal in many research scenarios. In order to provide a sound statistical basis for such comparative studies, there are various statistical tests for the detection of deregulated genes, proteins, or other molecular entities, which we will refer to as 'entity-level statistics'. Some of the most commonly used entity-level statistics will be described in the following paragraphs [312].
Consider the following notation: Let $x = (x_1, ..., x_n) \in \mathbb{R}^n$ and $y = (y_1, ..., y_m) \in \mathbb{R}^m$ be two series of measurements for a sample group and a control group, respectively. The sample mean $\bar{x}$ for a sample x of size n is defined as:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The sample variance s_x^2 for a sample x of size n with mean $\bar{x}$ is defined as:

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 .$$

The sample mean $\bar{x}$ ($\bar{y}$) and the sample variance s_x^2 (s_y^2) are estimators for the mean μ_x (μ_y) and variance σ_x^2 (σ_y^2) of the (respective) underlying population.

(Log) fold change: One of the simplest and most intuitive entity-level scores is the (log) fold change (sometimes called 'fold quotient'). The fold change is defined as the quotient between the means of the respective sample groups x and y:

$$\text{fc} = \frac{\bar{x}}{\bar{y}}$$

In order to adjust the ranges of scores indicating over- and underexpression, typically the log fold change is considered instead of the fold change:

$$\begin{aligned} \log_2(\text{fc}) &= \log_2 \left(\frac{\bar{x}}{\bar{y}} \right) \\ &= \log_2(\bar{x}) - \log_2(\bar{y}) \end{aligned}$$

By taking the logarithm of the original fold changes, the transformed scores are distributed around zero. While zero indicates that the expression levels remained constant between the two considered groups, positive scores indicate the upregulation of genes in the sample group in comparison to the control group, and negative scores indicate the downregulation of genes in the sample group in comparison to the control group. Besides its application on sample groups, the fold change can also be used to compare two samples directly, e.g. the gene expression of a specific gene in a tumor versus its expression in healthy tissue of the same patient.

Standard score: When comparing a single sample i (e.g., obtained from a specific tumor) against a larger background distribution (e.g., of a cohort of tumors of the same subtype), the

standard score (or z-score) is a typical measure of choice. The standard score is defined as

$$z_i = \frac{x_i - \bar{x}}{s_x},$$

where x_i denotes the expression value of the considered gene in sample i and the estimated mean and standard deviation of the background distribution are given by $\bar{x}$ and s_x, respectively. The z-score assumes normally distributed data and describes the number of standard deviations s_x the expression of sample i is away from the mean of the population $\bar{x}$.

T-tests: In scenarios where two series of (normally distributed) values should be compared, t-tests are typically applied. The t-test family consists of various hypothesis tests (cf. **Section 3.3.1**) that test for the inequality of means ($H_0 : \mu_X = \mu_Y$ vs. $H_1 : \mu_X \neq \mu_Y$). They can be applied in various scenarios (one-sample vs. two-sample, paired vs. unpaired) [313, 314]. One commonly used variant is the unpaired two-sample Student's t-test [315]. The values of its test statistic $t = T(x_1, ..., x_n, y_1, ...y_m)$ are distributed according to a t-distribution with $m + n - 2$ degrees of freedom and can be computed as:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}} \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

The Student's t-test assumes that the populations underlying the two samples have the same variance σ^2, which is approximated by their respective sample variances s_x^2 and s_y^2. There are also other variations of t-tests like the Welch's t-test [316, 317], which does not require the assumption of equal population variances. The Student's t-test is typically recommended for research scenarios with larger cohorts [313]. In cases with smaller sample sizes, regularized versions are preferred, such as the independent shrinkage t-test proposed by Opgen-Rhein and Strimmer [318].

Wilcoxon rank-sum test: The Wilcoxon rank-sum test [319] (also called the Wilcoxon-Mann-Whitney test) is a non-parametric alternative to the independent two-sample t-test, which is solely based on the relative order of the values in the two given samples (i.e., their ranks) and hence does not make any assumption about the underlying distributions.
The test statistic U is calculated based on the combined ranking R of the $n + m$ values in the two given samples x and y:

$$U = n \cdot m + \frac{n \cdot (n+1)}{2} - R_x,$$

where R_x is the sum of the ranks of those values in R that belong to the sample group x:

$$R_x = \sum_{i=1}^{n+m} r_i \cdot i,$$

with r_i being an indicator variable that is set to 1, if the i-th element in the ranked list R belongs to the sample group x, and to 0 otherwise. In order to assess the significance of an observed value of the test statistic, it can then be compared to the critical value for a given significance level α

and the corresponding samples size n and m as provided in a reference table (e.g., [314]). For larger sample sizes, the test statistic can be approximated by a normal distribution [320], based on which p-values can be derived.

3.3.3 Detecting deregulated pathways and networks

In the previous section, we presented several entity-level statistics that can be used to assess the degree of deregulation of genes, proteins, miRNAs, or other molecular entities. These methods might yield tremendously large lists of relevant, but isolated items. However, as already mentioned in **Section 2.1**, genes and their gene products do not act in isolation, but instead, they interact with each other in a highly coordinated and balanced fashion to form complex biochemical processes and signaling cascades that allow the cells to perform their designated functions. Given the functional interdependencies between the molecular components in a cell, complex diseases are rarely a consequence of an aberration in a single gene, but rather arise from the interplay of various causative factors and pathobiological processes. Hence, besides the investigation of single entities, the investigation of groups of genes in their functional context will provide a holistic view on a condition under investigation.

An important concept describing the complexity of cellular processes is the notion of a functional module, or pathway, as proposed by Hartwell *et al.* [321]. According to this simplified view, a module is a well-defined entity that is separable from other modules and whose components interact with each other to give rise to a specific biological function. While this view of biological processes and signaling cascades as separable units is favorable for various applications like the functional annotation of gene sets, one has to acknowledge that such pathways are only human-defined excerpts of large and highly interconnected molecular networks [19].

With the goal of providing biological context to high-throughput experimental data, a large variety of pathway and network analysis methods were proposed over the years. These approaches generally combine *a priori* knowledge of biological processes (cf. **Section 3.2.2**) with statistical and algorithmic procedures to elucidate complex biological mechanisms like the initiation and progression of cancer [322]. Following the classification proposed by Khatri *et al.* [323], we will present three major classes of pathway and network analysis approaches in the subsequent sections. The classes differ in the extent of *a priori* knowledge and types of input data used and range from Over-Representation Analysis (**Section 3.3.3.1**) over Functional Class Scoring (**Section 3.3.3.2**) to topology-based methods (**Section 3.3.3.3**). While most of the methods can be applied to various types of molecular entities, i.e. genes, proteins, miRNAs, and others, we will in the following only refer to 'genes' for the sake of readability. For each of these classes, several representative approaches and methods will be introduced.

3.3.3.1 Over-Representation Analysis

Life science experiments oftentimes yield a list of 'interesting' genes (e.g., a set of genes differentially expressed between two phenotypes). A typical follow-up question in such cases is whether the identified genes share a common biological function. To this end, Over-Representation Analysis (ORA) can be used. ORA tests for a set of interesting genes (called the 'test set') whether they are over- or under-represented in a so-called 'category' of genes with known functional annotation.

Consider an universe of possible entities, which is called the reference set R, a test set T, which is a subset of the universe ($T \subset R$), and a category $C \subset R$. The number of entities from the test set that overlap with the category $k := |T \cap C|$ can be compared with the number of entities expected to be found in this category just by chance using a classical urn model (cf. **Figure 3.4 A**): Given an urn containing $N := |R|$ balls out of which $K := |C|$ are red and $N - K$ are blue, and from which $n := |T|$ balls are randomly drawn without replacement. Furthermore, consider a random variable X that counts the number of red balls in a random sample. The probability of obtaining exactly k red balls in a random sample can be computed using the hypergeometric distribution [324]:

$$Pr(X = k \mid N, K, n) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

The expected number of hits in a randomly chosen test set of cardinality n is given by $k' = \frac{K \cdot n}{N}$. Based on the value of k', the hypergeometric test can be used to compute a one-sided p-value for the over-representation ($k > k'$) or under-representation ($k \leq k'$) of the category C in the test set T [325]:

$$p_C = \begin{cases} \sum_{i=k}^{n} Pr(X = i \mid N, K, n) & \text{if } k > k' \\ \sum_{i=0}^{k} Pr(X = i \mid N, K, n) & \text{if } k \leq k' \end{cases}$$

However, the above equations are only valid in cases where the test set is a subset of the reference set ($T \subset R$). If this is not the case, Fisher's Exact Test [300] should be used instead:

$$Pr(X = k \mid N, K, n, i) = \frac{\binom{n}{i}\binom{N}{K+k-i}}{\binom{N+n}{K+k}}$$

The p-values are defined analogously:

$$p_C = \begin{cases} \sum_{i=k}^{n} Pr(X = i \mid N, K, n, k) & \text{if } k > k' \\ \sum_{i=0}^{k} Pr(X = i \mid N, K, n, k) & \text{if } k \leq k' \end{cases}$$

As indicated above, ORA is a commonly used method to test for the functional enrichment of a sets of interesting genes as obtained from biological experiments. ORA can also be applied for the downstream analysis of high-throughput experiments that yield measurements for a large number of genes. However, in such cases, a subset of, for example, differentially expressed genes for further investigation has to be determined, which requires the user to select a threshold. Due to the arbitrary nature of such thresholds, potentially interesting components close to the cutoff threshold will be omitted, which might affect the results. Hence, in these cases, the methods presented in the following sections might be advantageous.

3.3.3.2 Functional Class Scoring

Similarly to Over-Representation Analysis, Functional Class Scoring (FCS) methods also aim at the identification of those pathways and functional gene sets that are enriched or depleted in the

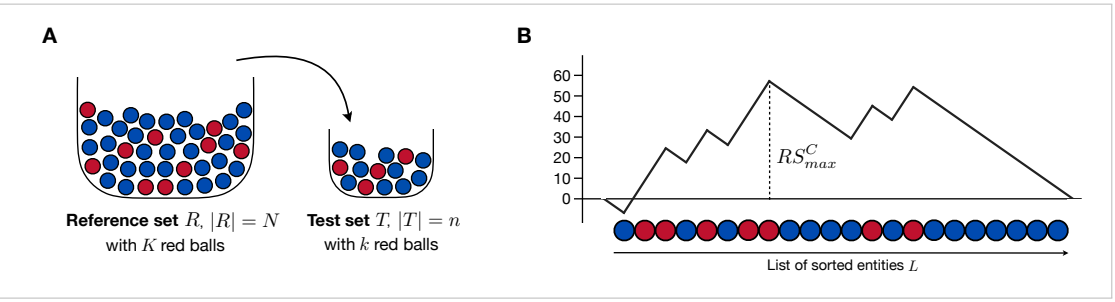


Figure 3.4 **Exemplary functional enrichment methods**. Entities belonging to a category of interest are colored in red, entities not belonging to the category in blue. **A) Urn model for the hypergeometric test.** A test set of size n is obtained from the reference set of size N via random sampling without replacement. **B) Exemplary Kolmogorov-Smirnov running sum as used in unweighted GSEA.** The entity list L is sorted based on the entities' associations with the considered phenotype. The maximum deviation of the running rum from zero (RS^C_{max}) serves as the value of the test statistic.

data set under investigation. However, in contrast to ORA, FCS approaches do not just consider sets of interesting genes, but take all measured genes as well as their scores of deregulation into account. In the following paragraphs, we will present the seminal method of Gene Set Enrichment Analysis as well as more recently proposed approaches.

Gene Set Enrichment Analysis: One of the first and most popular representatives of the FCS approach is Gene Set Enrichment Analysis (GSEA) [326, 327]. Based on scores of differential gene expression (cf. **Section 3.3.2**), GSEA ranks all measured entities (typically in decreasing order) and calculates a statistic that reflects the degree to which a given category is represented at the extremes of the ranked list, i.e. whether the category genes are enriched at the beginning or the end of the sorted list. This approach is formalized using the Kolmogorov-Smirnov running-sum statistic [328]: Consider a list of entities $L = (l_1, l_2, ..., l_n)$, sorted according to their entity-level scores. An example for this would be a list of genes in decreasing order of their t-scores indicating differential gene expression. The list L is traversed from top to bottom, while a statistic (the so-called 'running sum') is computed. The runnig sum starts at zero and is increased every time an entity belonging to a category C (with $m := |C|$) is encountered, and decreased otherwise:

$$RS^C(0) = 0$$
$$RS^C(i) = \begin{cases} RS^C(i-1) + w_i^+ & \text{if } l_i \in C \\ RS^C(i-1) - w_i^- & \text{if } l_i \notin C \end{cases}$$

The values of the respective increments (w_i^+) and decrements (w_i^-) depend on whether the original Kolmogorov-Smirnov formulation is used ('unweighted' GSEA) or its weighted extension as proposed by Subramanian *et al.* [326]. In both cases, the magnitudes of the respective summands are selected in such a way that the running sum returns to a value of zero after the traversal of the list L, i.e. $RS^C(n) = 0$. In the original, unweighted formulation, this is achieved by the two constants $w_i^+ = n - m$ and $w_i^- = m$. In weighted GSEA, for each entity

$i \in C$ with corresponding entity-level score $w(l_i)$, the respective increment w_i^+ is computed as:

$$w_i^+ = \frac{|w(l_i)|^p}{N_R},$$

where $N_R := \sum_{l_i \in C} |w(l_i)|^p$ is used as a normalization factor with the parameter $p \in \mathbb{R}_0^+$ controlling the degree to which the entity-level scores are used to weight the running sum. Conversely, the decrement is defined as:

$$w_i^- = \frac{1}{n-m}$$

As indicated in **Figure 3.4 B**, the enrichment score RS_{max}^C is defined as the maximum deviation from zero encountered during the traversal of the list:

$$RS_{max}^C = \max_i |RS^C(i)|$$

The enrichment score RS_{max}^C will become larger the more the entities contained in the considered category C tend to occur at the top or bottom of the given sorted list L.
The significance of an enrichment score RS_{max}^C can be determined using permutation tests. There are several approaches to this end, each of which has specific strengths and weaknesses, as well as data requirements, see Powers *et al.* [329] for an elaborate discussion. Performing sufficiently many permutation runs allows the generation of a null distribution, based on which an empirical p-value can be inferred. This p-value is based on the proportion of random permutation runs that obtained an enrichment score equal to or more extreme than the originally observed one.
In the case of unweighted GSEA, an algorithm for the calculation of exact p-values was proposed by Keller *et al.* [327], which has the advantage that the obtained p-values are not limited by the number of permutations conducted. In their method, the actual number p^* of running sum statistics that achieve an absolute score smaller than RS_{max}^C is determined using a dynamic programming approach. The exact p-value can then be computed as:

$$p = 1 - \frac{p^*}{\binom{n}{m}}$$

Finally, the obtained p-values are adjusted to account for multiple testing (cf. **Section 3.3.1.1**).
While the Kolmogorov-Smirnov-like statistics described above are very popular and widely used, there are a variety of other pathway-level statistics that are able to identify the coordinated up- or downregulation of entities belonging to a category of interest. The proposed statistical models range from simple statistics like the sum/mean/median gene level statistics [330] over the max-mean statistic [331] to the Wilcoxon rank-sum test (cf. **Section 3.3.2**), all of which have shown to yield comparable results to GSEA [332].

Pathifier: Similar to other FCS approaches, the Pathifier tool [333] aggregates entity-level information into pathway-level scores. For a given set of samples and pathways (categories) of interest, the method computes a Pathway Deregulation Score (PDS) for each sample-pathway pair. To this end, Pathifier requires a gene expression matrix $E \in \mathbb{R}^{p \times n}$ for p genes and n samples as input. Moreover, clinical or biological attributes (e.g., tumor aggressiveness or

patient survival) for the n samples need to be available. Finally, the sample set needs to contain several reference or control samples.

Each pathway analyzed in Pathifier is represented by a gene set P with $|P| = d_P$. For the analysis of a given pathway P, only the d_P genes constituting the pathway are considered for each of the n samples. Their respective gene expression scores ($E_P \in \mathbb{R}^{d_P \times n}$) are used to place all n samples in the d_P-dimensional subspace spanned by the pathway genes.

In a next step, a nonlinear 'principal curve' [334] is placed through the point cloud of samples following the progression of clinical or biological attributes of the samples (e.g., WHO grades, see **Figure 3.5 A** for illustration).

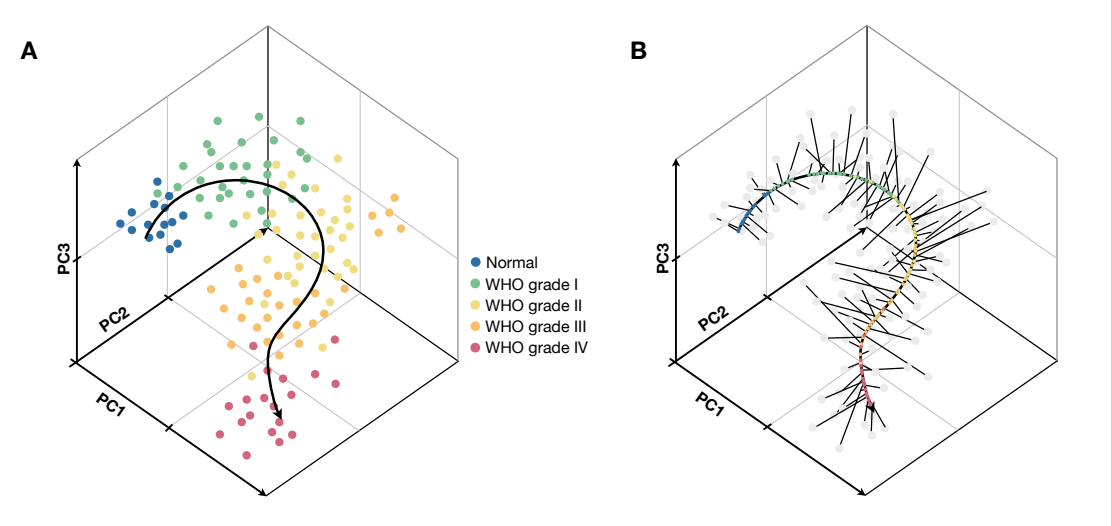


Figure 3.5 Exemplary visualization of the Pathifier approach. A) Principal curve following the progression of WHO grades. Principal component visualization of a point cloud of exemplary normal samples and tumor samples, colored according to their WHO grade. The principal curve is fitted following the progression of WHO grades from *Normal* over *WHO grade I* to *WHO grade IV* using the algorithm by Hastie and Stuetzle [334]. **B) Projection of samples onto the principal curve.** The samples are projected orthogonally onto the principal curve and the PDS of a specific sample i is estimated as the distance from the centroid of the normal samples (in blue), along the principal curve, to the projection of i. Figure based on [333].

In order to measure the PDS, the samples are projected orthogonally onto the principal curve (cf. **Figure 3.5 B**). The deregulation score $D_P(i)$ for a sample i is defined as the distance from the control state (i.e., the centroid of a set of control samples) to the sample's projection along the principal curve.

In contrast to GSEA, Pathifier does not provide p-values based on which the significance of single pathways can be assessed. Instead, the tool results in a matrix $M \in \mathbb{R}^{n \times m}$, with n being the number of samples and m the number of considered pathways, that contains a PDS for each pair. This matrix can then be used for clustering and hence the identification of potentially new, pathway-based molecular subtypes of cancer [335].

One major limitation of this approach is its reliance on a large cohort of samples, preferably with a set of corresponding healthy samples and comprehensive and meaningful clinical annotations, which is feasible for scenarios like the retrospective analysis of large data sets like TCGA (cf. **Section 3.2.3**), but impractical for the analysis of individual samples (e.g., for treatment stratification purposes).

Metagene-based methods: In the subclass of metagene-based methods, genes belonging to a pathway of interest (or a subset thereof) and their respective scores of deregulation are aggregated into the deregulation score of a 'metagene', which then serves as a proxy for the considered pathway's activity.

One of the first approaches in this realm was proposed by Guo *et al.* [336], who summarized the scores of differential expression of the pathway genes based on simple statistics like the mean or median. An alternative approach, presented by Bild *et al.* [337], considers for each pathway P the d_P-dimensional space spanned by the d_P genes of the pathway and performs a principal component analysis [338] on the embedded point cloud of samples. The pathway activation of a sample i then is represented by its first principal component.

In contrast to the former two methods that consider all pathway genes, Lee and coworkers only consider a subset of pathway genes for pathway activity inference. To this end, they propose pathway-specific sets of COndition-Responsive Genes (CORGs) [339]. A CORG set G_P^k is a subset of those k genes of a pathway P whose averaged differential expression delivers optimal discriminative power for the distinction of the two phenotypes of interest. **Figure 3.6** provides a summary of the approach.

In a first step, the differential expression of each pathway gene is computed using a t-test (cf. **Section 3.3.2**) comparing the two phenotypes under investigation. Depending on the average t-score among all pathway genes, either only genes with positive t-scores (for $\bar{t} \geq 0$) or only genes with negative t-scores (for $\bar{t} < 0$) are considered for the CORG set. Starting with an initial CORG set G_P^1 that only contains the gene with the largest differential expression (from the respective subset), the CORG set is iteratively extended by the next most differentially expressed gene ($G_P^k \rightarrow G_P^{k+1}$), until the discriminative power of the CORG set does not improve any more. In order to assess the discriminative power of a given CORG set G_P^k, the expression scores of the genes in G_P^k are averaged to obtain a pathway activity score for each sample i:

$$a_{P,i}^k = \frac{1}{\sqrt{k}} \sum_{j=1}^{k} e_{j,i}$$

with $e_{j,i}$ being the gene expression of gene j in sample i. The activity scores for the samples within the two phenotype groups are compared using a t-test to determine the discriminative power of the current CORG set G_P^k. The final result for a pathway P is the CORG set G_P^k with the smallest k with $1 \leq k \leq d_P$ satisfying $T(G_P^{k+1}) \leq T(G_P^k)$ and $T(X)$ being the t-test for the pathway activities computed based on the gene set X.

One central limitation of this approach is the focus on either upregulated or downregulated genes within a pathway, which, however, both contribute to the overall pathway's activity and hence should be investigated jointly.

Based on the CORG approach, Sootanan *et al.* try to address this problem using Negatively Correlated Feature Sets (NCFS). NCFS follows the CORG method in the sense that in a greedy fashion genes are iteratively added to a predictive set until the discriminative power converges. In NCFS, the pathway genes are split into two sorted lists, each ordered by the genes' absolute scores of correlation or anti-correlation, respectively. **Figure 3.7** provides an overview of the approach.

Here, the predictive set constitutes of two subsets $G_P^{k_1}$ and $G_P^{k_2}$ that are iteratively extended in each round, each by the feature with the next largest correlation and anti-correlation,

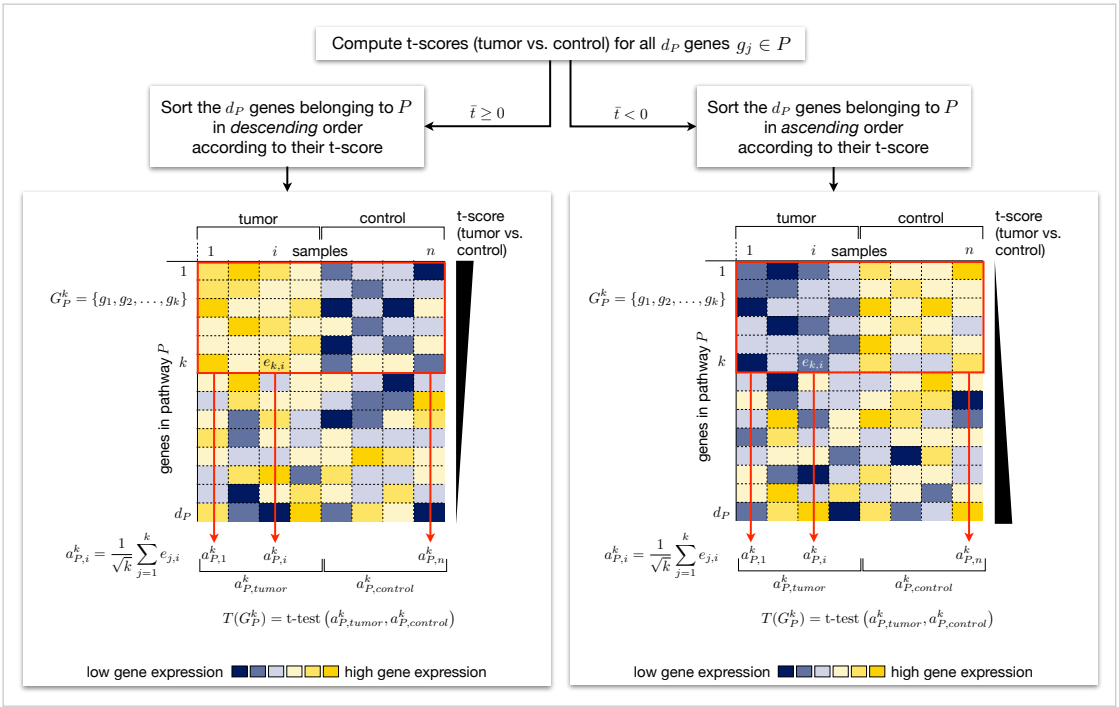


Figure 3.6 Schematic overview of the CORG approach. For all d_P genes in a pathway P, t-scores assessing the differential gene expression between two phenotypes (e.g., tumor vs. control) are computed. Depending on the average t-score $\bar{t}$ across the d_P genes, two different cases are considered. The CORG set G_P^k (highlighted in red) is iteratively extended and its discriminative power with respect to the two considered phenotypes is determined using a t-test ($T(G_P^k)$) on their respective pathway activities $a_{P,tumor}^k$ and $a_{P,control}^k$. Figure based on [339].

respectively. The pathway activity $a_{P,i}^k$ of a pathway P in a sample i is then computed as the difference of the averaged scores of genes in the positively correlated set $G_P^{k_1}$ and the averaged scores of genes in the negatively correlated set $G_P^{k_2}$:

$$a_{P,i}^k = a_{P,i}^{k_1} - a_{P,i}^{k_2}$$

The discriminative power of the predictive set is iteratively assessed analogously to the CORG method described above.

Although FCS-based approaches successfully alleviate the problems of ORA, they still face some limitations that arise from the consideration of pathways as mere sets of genes. Most importantly, FCS methods do not consider the dependencies between genes in a pathway, neither their relative positions within the pathway. Also, FCS approaches are limited by the fact that each gene set is analyzed individually, not accounting for the strong interdependence and cross talk between pathways at a systems level.

3.3.3.3 Topology-based methods

Over the last decades, enormous efforts have been put into the investigation of molecular interactions, many of which are described in various large-scale databases (cf. **Section 3.2.2**). Computational approaches that take these interactions into account are of particular interest, especially for the detection of pathways and subnetworks that are deregulated in pathogenic

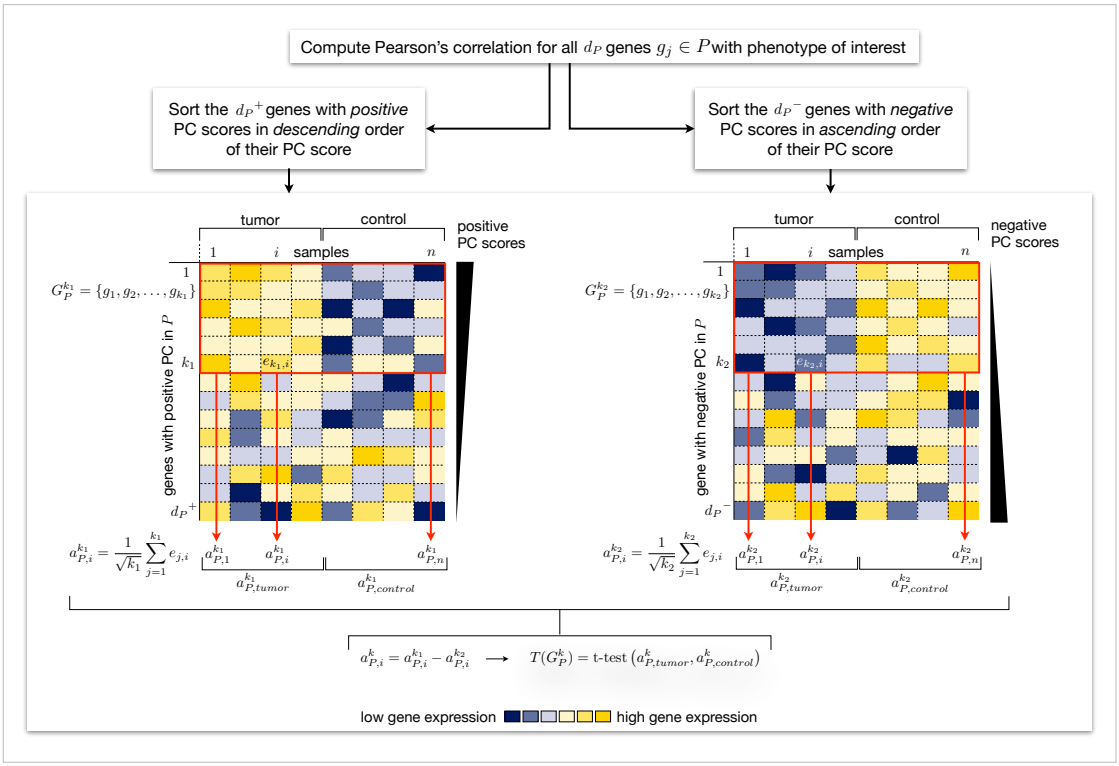


Figure 3.7 Schematic overview of the NCFS approach. For all d_P genes in a pathway P, their Pearson's correlations (PC) with the marker of interest are computed (e.g., survival time, drug sensitivity). Genes with positive correlation are sorted in descending order of their correlation score and genes with negative correlation are sorted in ascending order. The predictive set $G_P^k = G_P^{k_1} \cup G_P^{k_2}$ (highlighted in red) is iteratively extended and its discriminative power with respect to the two considered phenotypes is determined using a t-test ($T(G_P^k)$) on their respective pathway activities $a_{P,tumor}^k$ and $a_{P,control}^k$. Figure based on [340].

processes. This third class of methods combines information about the deregulation of genes between two phenotypes with topology information on individual pathways or their integration into networks.

Prior to the presentation of various topology-based methods, we first have to consider the mathematical representation of biological networks, which is common to all of these methods. In general, a biological network can be modeled as a graph $G = (V, E)$, whose nodes $v_i \in V$ represent biological entities, such as proteins, functional RNA molecules, or metabolites, while the edges $e_{ij} = (v_i, v_j) \in E$, $v_i, v_j \in V$ correspond to relationships between those biological entities, for example, binding, activation, or inhibition. The network interactions can be modeled either by undirected or by directed edges, representing, for example, protein-protein interactions or activating/inhibitory regulatory events, respectively. Depending on the type of entity and interactions considered, various types of biological networks can be distinguished [341]: Protein-protein interaction networks contain proteins as nodes and (typically) undirected edges that indicate an interaction of two or more proteins. Metabolic networks encode the biochemical reactions of metabolites being catalyzed by enzymes. Gene regulatory networks describe how genes regulate each other, and in a related manner, signaling networks trace the information flow in and between cells.

The continuously growing class of topology-based methods includes various sub-classes that differ in their computational approaches as well as their biological goals: some methods aim at predicting the activities of pre-defined pathways, similarly to FCS approaches, while others focus on the identification of deregulated subnetworks and functional modules. In the following paragraphs, we will exemplarily outline several methods from different sub-classes.

ScorePAGE: For the identification of active (metabolic) pathways, Rahnenführer *et al.* proposed the ScorePAGE algorithm (Scoring Pathway Activity with Gene Expression Data) [342]. ScorePAGE requires gene expression measurements of n samples under different conditions (e.g., time points after a perturbation) and a set of (metabolic) pathways, including their associated genes and respective topologies. For a pathway P under investigation, the ScorePAGE method computes for each pair of genes (g_i, g_j) within the pathway a similarity score $s(g_i, g_j)$. To this end, different measures like Pearson's correlation [343], covariance [344], or the cosine similarity [345] are used on the n samples. The similarity scores are moreover weighted using the distances between the two genes in the pathway, i.e. the minimal number of reactions connecting the two enzymes in a metabolic pathway.

The overall pathway activity $s(P)$ for a pathway P of size k is then given by the average of all considered gene pairs:

$$s(P) = \frac{1}{\binom{k}{2}} \sum_{1 \le i < j \le k} \frac{1}{\min\{d(g_i, g_j), 10\}} \cdot s(g_i, g_j),$$

where $d(g_i, g_j)$ is the distance between two genes g_i and g_j and the constant 10 is used to ensure a minimal contribution of all gene pairs within the pathway to the overall activity score.

The significance of the co-regulation of the genes within a pathway and hence its activity is assessed using a nonparametric permutation test. In order to account for multiple hypothesis testing, the Benjamini and Hochberg correction is performed (cf. **Section 3.3.1.1**).

SPIA: The Signaling Pathway Impact Analysis (SPIA) [346] aims at identifying deregulated molecular pathways via the combination of two types of evidence: (i) the overrepresentation of differentially expressed genes in the pathway of interest and (ii) the investigation of the pathway's overall 'perturbation' (i.e., deregulation). While the first part can be achieved using a classical ORA (cf. **Section 3.3.3.1**), the second part uses the network topology to trace the propagation of expression changes. To this end, for each gene g_i, its score of differential expression and the amount of perturbation in its downstream genes are taken into account. To this end, SPIA scores a gene highly 'perturbatory' if it affects other perturbatory genes in the network, see **Figure 3.8**. The perturbation factor $PF(g_i)$ of a gene g_i is computed using a recursive algorithm similar to the PageRank index used by Google [347]:

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

Here, $\Delta E(g_i)$ is the score of differential expression of gene g_i. The second term is the sum of the perturbation factors $PF(g_j)$ for all n genes g_j directly upstream of the target gene g_i, normalized by the number of all genes directly downstream of each such gene $N_{ds}(g_j)$, and weighted by β_{ji}. The indicator variable β_{ji} reflects the type of interaction present between two network nodes g_j and g_i, with $\beta_{ji} = 1$ indicating an activating effect and $\beta_{ji} = -1$ for an inhibitory effect of g_j on g_i. By following this recursive approach, the authors take into account the location of a gene in the pathway, following the rationale that deregulated genes at the beginning of a pathway can also perturb many downstream genes, while aberrations towards the end of a signaling cascade only affect fewer downstream genes.

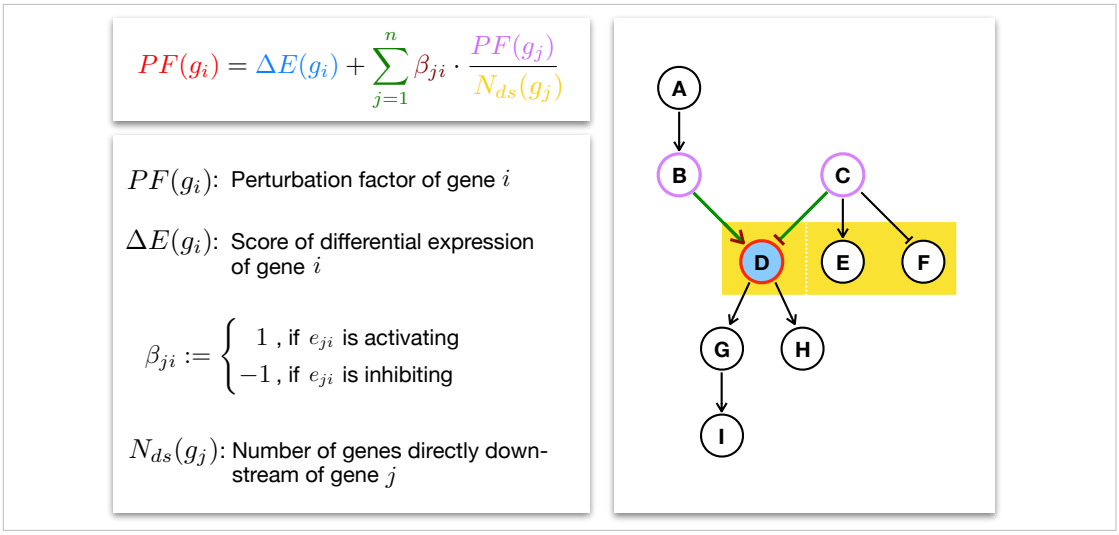


Figure 3.8 Perturbation factor computation in SPIA. The coloring in the exemplary pathway on the right visualizes the network components considered when computing the perturbation factor $PF(g_i)$ for a gene i, which corresponds to node D in the visualization. Pointed arrows indicate an activating interaction, whereas 'T'-shaped arrows indicate an inhibitory interaction.

Although this approach seems to be appealing, it has the major drawback that it considers pathways independently, neglecting their crosstalk and interdependencies. In order to alleviate this limitation, the authors recently proposed System-level PATHway Impact AnaLysis (SPATIAL) [348], an extension of SPIA that takes a global view on the pathway and network

topology and also considers the perturbation of respective upstream pathways in its pathway activity model.

Network diffusion-based methods: Another set of tools employs network diffusion-based approaches to study the effects of molecular aberrations (e.g., mutations) on downstream genes and processes. The HotNet [349] algorithm uses the physics of heat diffusion to detect 'mutated subnetworks', i.e. subnetworks of genes likely to be affected by the mutations present in a (tumor) sample. To this end, the network is considered as a metallic lattice and the genomic alterations as sources of 'heat' which propagates along the network topology, leading to 'hot', i.e. highly relevant, networks. The heat diffusion is here modelled as a random walk on the network graph. A random walk is a stochastic process that models the iterative transition of a 'random walker' from seed nodes (the affected, e.g., mutated, genes) to randomly chosen neighbors over time until a steady state is reached [350, 351].

EnrichNet [352] is another approach that uses random walks to assess the effect of aberrations on pathway activities. Starting from a set of seed genes, a network diffusion is performed. The diffusion scores obtained by the genes of a specific gene set (i.e., pathway or category) are then converted to distances, resulting in a distance vector for each pathway. This distance vector is compared to the average distribution across all pathways to assess the significance of the pathway's deregulation.

There are also several methods that extend the uni-directional network diffusion approaches described above: TieDIE (Tied Diffusion Through Interacting Events) [353] and NetICS (Network-based Integration of Multi-omiCS Data) [354] use directed graphs to perform a bidirectional graph diffusion from two different sources, namely a set of genes with genomic aberrations and a set of differentially expressed genes. While the 'heat' from genomic aberrations is diffused along the edges of the network, the scores of differentially expressed genes are diffused in the opposite direction, along reversed edges. By this, linker (or 'mediator') genes that connect genomic aberrations and transcriptional changes can be identified. These mediators might be promising candidates for the development of targeted therapies. TieDIE and NetICS follow the same analysis steps, where NetICS performs an additional step of averaging over the samples of a population to obtain a population-wide view on potential mediators.

Formulation as an optimization problem: Instead of focusing on specific seed genes as done by the presented diffusion-based approaches or the investigation of individual pathways as in SPIA, there is another class of methods that aims at identifying deregulated functional modules in biological networks.

To this end, these approaches employ the topology of a biological network G and scores of differential expression that are mapped onto the nodes of the network.

For example, Keller *et al.* proposed the Finding Deregulated Paths (FiDePa) algorithm [355], an algorithm that efficiently searches for all paths of length k in a given regulatory or signaling network that are significantly enriched with deregulated genes or proteins. To this end, paths of length k in the network are used as categories in unweighted GSEAs (cf. **Section 3.3.3.2**). In order to identify the most significant paths efficiently, a dynamic programming scheme was devised.

Other approaches aim at solving the Maximum-Weight Connected Subgraph (MWCS) problem, which is based on the idea of 'active subnetworks' as introduced by Ideker *et al.* [356]: Given a connected, undirected, vertex-weighted graph $G = (V, E, w)$ with weights $w : V \to \mathbb{R}$, find a connected subgraph $G' = (V', E') \subseteq G$ that maximizes the score $w(G') = \sum_{v \in V'} w(v)$. The MWCS problem has been proven to be NP-hard [356]. One of the first methods to approach the MWCS problem was presented by Ulitsky *et al.* with DysrEgulated Gene Set Analysis via Subnetworks (DEGAS) [357]. DEGAS searches for a minimal connected subnetwork with at least k differentially expressed nodes (i.e., genes or proteins) in all but l investigated samples. To solve this NP-hard problem, DEGAS employs heuristics with provable performance guarantees.
Another tool that solves a variation of this problem is KeyPathwayMiner [358]. KeyPathwayMiner employs an ant colony optimization technique to identify connected subnetworks of maximal size in which all but k nodes are differentially expressed in all but l analyzed samples.
One of the first approaches to solve the original MWCS problem to optimality was introduced by Dittrich *et al.* [359]. The authors reformulated the problem as a Prize-Collecting Steiner Tree problem, which can be efficiently solved using a corresponding Integer Linear Programming (ILP) formulation [360]. (Integer) Linear Programs are optimization problems with linear objective functions that are optimized with respect to linear constraints [361]. In the case of Integer Linear Programs, the results are constrained to be integer.
Backes *et al.* extended the approach by Dittrich and coworkers to account for directed networks, in order to better model the signal propagation within biological networks [362]. To this end, they proposed a branch-and-cut algorithm based on an ILP formulation.
The corresponding problem can be formulated as follows: Given a directed graph $G = (V, E)$, find the most deregulated subgraph G' of given size k where all nodes $v' \in G'$ are reachable from a designated root node v_R that also belongs to G'. This root node could be a molecular key player contributing to the deregulation of its downstream components and hence might be of interest as a potential drug target.
In the following, we will describe the corresponding ILP formulation in more detail, as this 'Subgraph ILP' will be further employed in the tools and analyses presented in the remainder of this thesis. To this end, we introduce the following notation: Let $w = (w_1, ..., w_n) \in \mathbb{R}^n_+$ be a vector of gene weights, containing for each node $i \in V$ its absolute score of differential expression. The binary vector $x = (x_1, ..., x_n) \in \mathbb{B}^n$ contains an indicator variable for each node of the graph, indicating the selection of subgraph vertices in the result ($x_i = 1$ if v_i is selected, 0 otherwise). Finally, the binary vector $y = (y_1, ..., y_n) \in \mathbb{B}^n$ contains indicator variables for the choice of the root node ($y_i = 1$ if node v_i is selected as the root node, 0 otherwise).
An overview of the problem formulation is given in **Table 3.1**: In order to find the most deregulated subgraph, the objective function is the sum of the selected subgraph nodes, weighted by their score of deregulation (**Equation 3.1**). **Equations 3.2** and **3.3** ensure that precisely k vertices are selected to be part of the solution and that exactly one root node is assigned. The **Inequations 3.4** and **3.5** guarantee for each selected vertex that it is either the root node or that at least one of its parent nodes (parents of i are given by the set $In(i)$) has been selected. The last constraint (**Inequation 3.6**) prevents the ILP to yield two disconnected cycles by making sure that each selected cycle C either contains the root node or that a selected vertex outside of C is a parent of one of the cycle nodes.

Objective		
$\max_{x \in \mathbb{B}^n} \sum_i w_i \cdot x_i$	(3.1)	Maximize the overall deregulation of the subgraph
Subject to		
$\sum_i x_i = k$	(3.2)	Ensures that the subgraph is of size k
$\sum_i y_i = 1$	(3.3)	Ensures that a single root node is selected
$y_i \leq x_i \quad \forall i$	(3.4)	Ensures that the designated root node is part of the selected subgraph
$x_i - y_i - \sum_{j \in In(i)} x_j \leq 0 \quad \forall i$	(3.5)	Ensures connectivity of the subgraph
$\sum_{i \in C}(x_i - y_i) - \sum_{j \in In(C)} x_j \leq \lvert C \rvert - 1 \quad \forall C$	(3.6)	Prevents disconnected cycles

Table 3.1 **Subgraph ILP formulation**. The objective function and the respective constraints are given in the first column, the second column provides a numbering for reference in the text, and the third column describes the purpose of the respective formula. The variable C describes cycles formed by the selected nodes.

A limitation of topology-based methods is their dependence on just this type of *a priori* knowledge. Molecular interaction networks are still not complete [363]. Besides a variety of missing links, both experimental techniques (e.g., yeast-two-hybrid) and computational inference approaches are prone to false positives [364, 365]. Moreover, currently available networks also are of rather coarse granularity as they typically do not distinguish between different splicing variants or isoforms of a gene or protein.

4

Tools for Multi-Omics Integrative Analyses

The thorough measurement of biological data using, for example, the experimental high-throughput techniques described in **Section 3.1**, allows investigating biological systems at an unprecedented depth and scale. In combination with additional resources of *a priori* knowledge (cf. **Section 3.2**), this enables a holistic view on the mechanisms underlying these biological systems, and thereby provides means to analyze the emergence and progression of complex diseases like cancer. The processing, handling, integration, annotation, and analysis of such complex, heterogeneous, and often large data sets calls for a computational infrastructure and novel methods to generate a multi-dimensional picture of the conditions under investigation. To this end, we have developed an integrative tool suite for computational systems biology that facilitates systems-oriented research by providing various tools and methods for multi-omics integrative analyses and biomarker identification, which will be presented in this chapter. First, we will present Graviton, a general framework for the implementation of web-based, integrative, multi-omics systems-biology tools. Graviton also serves as basis for our specialized analysis pipelines (**Section 4.1**): GeneTrail2 - a web service for multi-omics enrichment analysis (**Section 4.2**), RegulatorTrail - a web service for the identification of key transcriptional regulators (**Section 4.3**), and NetworkTrail - a web service for identifying and visualizing deregulated biological subnetworks (**Section 4.4**). Two specialized analysis tools for personalized medicine that focus on drug repositioning and clinical decision support will be presented separately in **Chapters 5** and **6**.

4.1 Graviton - a framework for multi-omics integrative analyses

> The framework has mainly been developed by Daniel Stöckel and Tim Kehl. I have contributed to Graviton via the integration of additional databases, tools, and analyses that enable the use of Graviton in the context of personalized medicine.

The main challenge in systems biology is the complexity of the investigated biological systems in combination with the vast amount of data that is scattered across numerous resources and that all have to be integrated and jointly analyzed. Hence, comprehensive computational tools are required to gain novel insights in systems biology [366, 367]. The integration and combined analysis of data from different sources is a multistep procedure involving a variety of tools for data handling and harmonization, the actual analyses, the visualization of the results, and their export for downstream processing.

To provide a sound basis for tackling these challenges, we developed Graviton, a general framework for integrative multi-omics systems-biology approaches. Graviton provides a wide array of general-purpose functionality like data upload, identifier mapping, resource handling, job scheduling and forms the basis for all of our specialized tools and web services, which will be discussed in subsequent sections. First, we will briefly describe the enabling potential of such a framework (**Section 4.1.1**), provide an overview of its implementation (**Section 4.1.2**), and describe its workflow (**Section 4.1.3**).

4.1.1 Software as a service

Software as a Service (SaaS) is a software distribution model in which a provider hosts an application and customers access the offered functionality over the internet, typically using standard web browsers [368]. While SaaS is mainly used in commercial context, it also provides numerous advantages for academic settings, and especially for the use in systems biology approaches: The implementation of systems biology research projects requires the cooperation of researchers from various domains (such as biologists, physicians, mathematicians, chemists), not all of which necessarily have a background in bioinformatics or computer science. However, for the integrative analysis of multi-omics data, the installation and correct use of various tools are required, which typically have to be operated over the command-line and do not provide a Graphical User Interface (GUI) [369]. Furthermore, although there are many standards for data representation, the output and input formats of different tools are commonly not interoperable and require reformatting [370]. This impedes productivity and is prone to errors. The implementation of analysis pipelines in the form of web services overcomes these obstacles: A web service only requires a single, centralized point of installation, which is maintained by the host. The provided functionality is platform-independent and accessible from anywhere. Moreover, the seamless integration of the respective analysis steps mitigates the otherwise tedious task of individually performing all steps of the analysis with specialized tools or scripts. There are various workflow-management systems that aim at providing such integrated functionality. Systems like KNIME [371], Taverna [372], or Galaxy [373] enable the construction and execution of specialized workflows. While these platforms allow for convenient and customized pipeline design, they do not provide ready-to-use workflows, but instead require the user to select and arrange the individual analysis steps from a multitude of options.

4.1.2 Implementation

Graviton is implemented based on a modular multi-layer client-server architecture that ensures extensibility and maintainability (cf. **Figure 4.1**).
The first layer corresponds to the user frontend, which is implemented using HTML5 [374] and CSS3 [375], in combination with Bootstrap [376] and the Thymeleaf template engine [377]. Results are visualized using the DataTables plug-in for JQuery [378] and the JavaScript libraries of Chart.js [379], D3.js [380], and Highcharts [381]. In the application layer, Java [382], JavaScript [383], JQuery [384], and AJAX [385] are used for the implementation of the internal logic and client-server communication with a RESTful Application Programming Interface (API) [386]. The API allows to set up and run computationally intensive tasks ('jobs') and to handle the created (intermediate) results ('resources'). The last layer, the resource layer, contains rich analysis

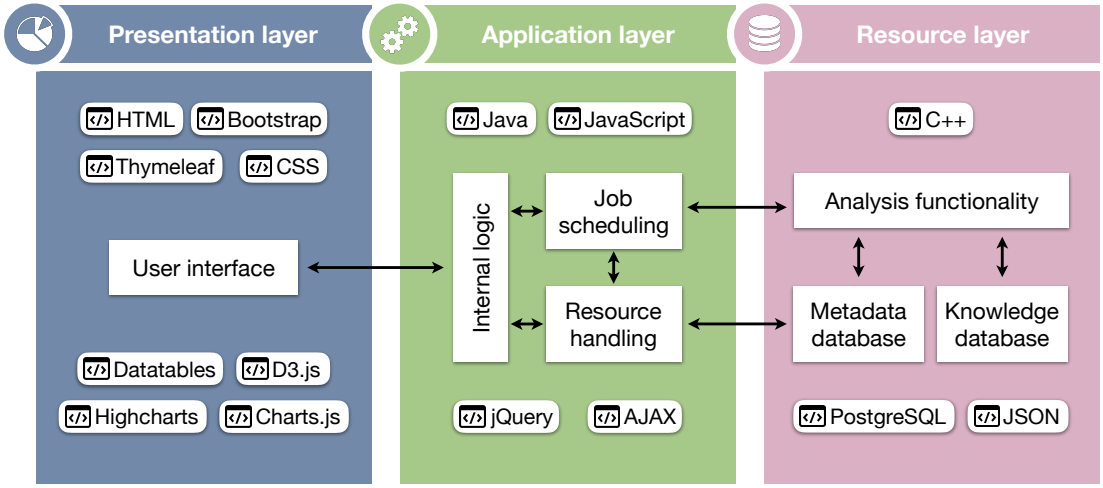


Figure 4.1 **Graviton architecture**. The colored boxes represent the layers of the three-tier client-server architecture. Within each layer, the respective components and employed technologies (indicated by the 'code' icon) are displayed. The displayed icons were obtained from [15].

functionality implemented in the form of a C++ library [387]. Metadata on user sessions and performed analyses are stored in a PostgreSQL database [388]. This ensures the reproducibility of the obtained results, as all analysis parameters are recorded, including the random seeds that are used in analysis steps involving randomization. Furthermore, we use a document-oriented database [389] for the storage of *a priori* biological knowledge, which is used for annotation and analysis of the uploaded data sets. The underlying database content is regularly updated in a semi-automatized manner.

Documentation: In order to make a tool or web service as easy to use as possible for researchers of various domains, an intuitive user interface and thorough documentation are essential. To this end, the user-friendly interface of our web services leads the user through the data upload and analysis steps and provides reasonable default parameter settings, matching the properties of the uploaded data. Moreover, we provide extensive documentation for all our web services. The supporting information ranges from standalone tutorials for the respective analysis workflows, over example files and additional information along the data upload and analysis steps, to interactive explanations of the provided results. All visualizations and (intermediate) results can be downloaded for further processing or reporting.

RESTful API: Graviton offers a large variety of tools and methods, which are useful in a variety of research scenarios. In order to enable other bioinformaticians to integrate this functionality into their existing analysis pipelines, Graviton also provides a RESTful API as a programming interface. A RESTful API is an API that follows the architectural style of REpresentational State Transfer (REST). REST is a scheme for client-server communication that allows the client to access and modify textual representations of Web resources using a predefined set of stateless operations [390]. Web resources can be identified over the internet via their Uniform Resource Identifier (URI). Typical operations in REST are `GET`, `POST`, `PUT`, or `DELETE`. These operations are called 'stateless' because a client's request is required to contain all the

information (parameters) necessary to be processed by the server and hence no contextual information needs to be stored on the server. An overview of the Graviton API is provided on `http://apidoc.bioinf.uni-sb.de` and an example of how to programmatically run an analysis is listed in **Section A.3.1**.

4.1.3 Workflow and functionality

Graviton is a framework for building fully integrated bioinformatics web services. To this end, it provides implementations of general-purpose functionality like file parsers, scoring, identifier mapping, and annotation of data sets, which can be used in all of our specialized web services and hence avoids code duplication and reimplementation of functionality. **Figure 4.2** gives an overview of the Graviton workflow.

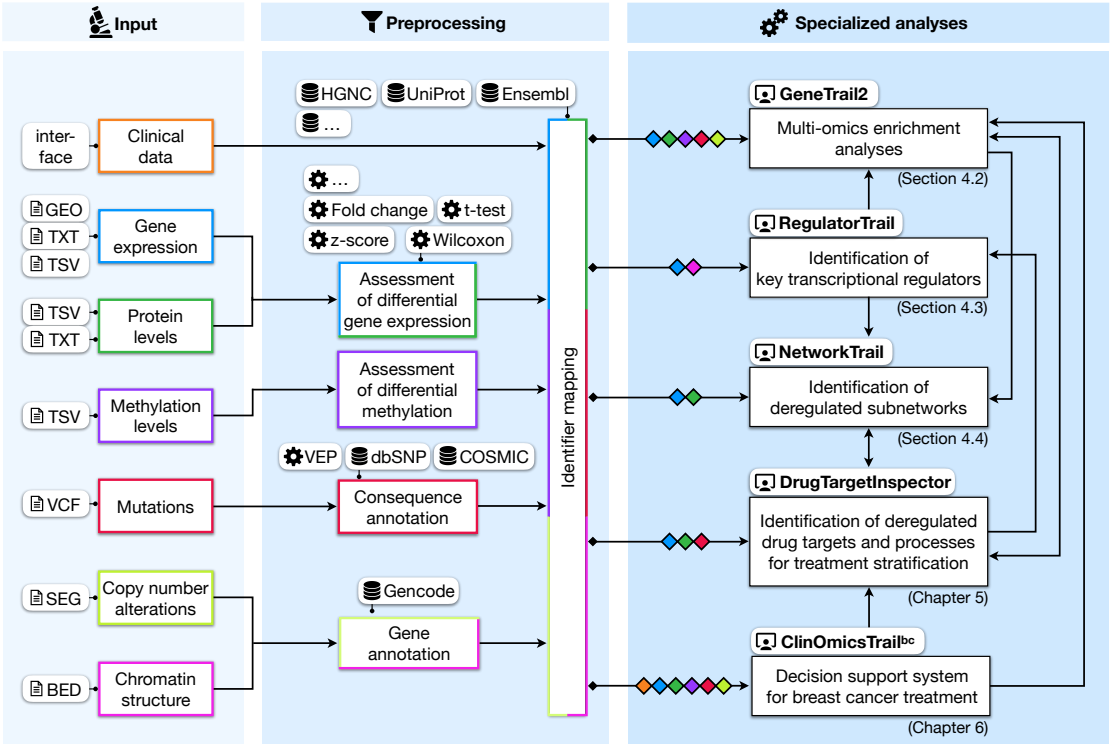


Figure 4.2 **Overview of Graviton workflow**. The box border and diamond colors correspond to the type of data used in the corresponding step or analysis. Orange: clinical data, blue: gene expression data, green: protein levels, purple: methylation data, red: mutation data, lime: copy number alterations, pink: chromatin structure. Input data formats are indicated by the 'file' icon, used databases by a 'database' icon, statistical and computational methods by a gear wheel, and our integrated web services by the 'user' icon. The section and chapter numbers in the 'specialized analyses' box are references to the sections and chapters in which the respective tools are presented. BED: Browser Extensible Data, COSMIC: Catalogue Of Somatic Mutations In Cancer, GEO: Gene Expression Omnibus, HGNC: HUGO Gene Nomenclature Committee, SEG: SEGmented data file format, TSV: Tab-Separated Values, TXT: plain text file, VCF: Variant Call Format, VEP: Variant Effect Predictor. The displayed icons were obtained from [15].

Supported input data: Graviton provides the basis for several specialized multi-omics integrative web services and hence supports a large variety of data input types. The framework enables the upload of numerous types of omics data in several (standardized) input formats (cf. **Figure 4.2**, first column). For the use in gene set based analyses (e.g., as provided by GeneTrail2

or RegulatorTrail), a plain text file containing a list of identifiers of a set of 'interesting entities' can be provided as most simple form of input. Transcriptomics, proteomics, and epigenomics data can be uploaded as whitespace-separated files (e.g., in `TSV` file format) that contain a whitespace-separated pair of entity identifier and corresponding deregulation score per line. Besides files containing user-defined scores, also matrices containing expression values for a set of samples can be uploaded. Alternatively, expression matrices from GEO [294] can be imported by providing identifiers for GEO SEries (`GSE`) or GEO Data Set (`GDS`) files. The samples contained in such matrices can be divided by the user into the samples under investigation and a set of control samples. The user can then analyze the two groups using various types of entity-level statistics, as described in the following paragraph. Mutation data can be uploaded in Variant Call Format (`VCF`, cf. **Section A.1.4**) and should contain information on a tumor sample and ideally a matched control to be able to differentiate between somatic and germline mutations. Copy number alterations can be provided to the web service in SEGmented data file format (`SEG`, cf. **Section A.1.6**). Besides providing flexibility in the data formats, we also aim at covering as many of the standard entity identifier types as possible. These include EntrezGene [270], Ensembl identifiers [273], HGNC symbols and IDs [272], KEGG [282], UniProt [274] for genes and proteins, and miRBase [278] identifiers for miRNAs.

Preprocessing: Depending on the type of uploaded data, up to three additional preprocessing steps are performed (cf. **Figure 4.2**, second column) before the data can serve as input for the specialized analyses. The preprocessing steps include (i) the assessment of differential gene expression or methylation levels, (ii) the annotation of genomic aberrations with additional metadata, and (iii) the harmonization of the respective entity identifier types used in the uploaded files.

In cases where a matrix of measurements for samples of two phenotypes (e.g., tumor and control) is provided, the user can select for both phenotypes those samples that should be used in the analysis. In a next step, the groups of selected samples are compared against each other to assess differential gene expression or methylation. Depending on the sample selection, the user can choose from a broad array of parametric and nonparametric tests and scoring schemes (cf. also **Section 3.3.2**). In cases where one sample of interest is compared against one or several reference samples, the (log) mean fold quotient [391] can be computed. If a reference set with several samples is provided, the user could additionally select the standard score (z-score) [392] as the scoring method. For the comparison of two groups of samples, the following tests are provided: independent shrinkage *t*-test [393], independent Student's *t*-test [394], Wilcoxon-Mann-Whitney test [395], signal-to-noise ratio [396], *F*-test [397], and the (log) mean fold quotient [391]. A complete list of all provided entity-level statistics is given in **Table A.4**.

If genomic aberrations in the form of mutations are uploaded, Graviton annotates the mutations with their predicted impact on the corresponding genes and proteins (e.g., missense variant, stop gain, frameshift). To this end, Ensembl's Variant Effect Predictor (VEP) [220] is used (cf. **Section 3.1.2.3**). Additionally, the contained mutations are cross-referenced with dbSNP [226] and COSMIC [225] for further details on the potential functional impact and pathogenicity of the mutation. Since copy number alterations are typically described in terms of the genomic location (coordinates) they occur in, the affected genes have to be identified using reference genomes (e.g., GRCh37/38 [398]) and gene annotations, which we obtained from Gencode [399].

The integration and analysis of heterogeneous data from different sources requires their harmonization into a common format using common identifiers. To this end, Graviton provides sophisticated mapping functionality [400] and mapping files for a variety of organisms and a plethora of identifier types (e.g., from NCBI EntrezGene [270], Ensembl [273], HGNC [272], KEGG [282], UniProt [274], miRBase [278]). A complete list of all available mappings is given at `https://genetrail2.bioinf.uni-sb.de/mappings.html`. As the internal representation of gene- and protein-based entities within Graviton, HUGO gene symbols are used (cf. **Section 3.2.1**). In order to keep the mapping process traceable and transparent, Graviton provides mapping statistics that indicate which identifiers were mapped to which other identifiers and also which identifiers could not be mapped and hence were not considered in the further analysis steps.

Specialized analyzes: After upload and preprocessing of the different omics data sets, the data can be analyzed using various specialized tools and analysis pipelines, which will be presented in detail in the following sections and chapters (cf. **Figure 4.2**, third column). GeneTrail2 is a web-interface providing access to different tools for the statistical analysis of molecular signatures with a focus on pathway enrichment analyses. It offers multiple statistical tests and a comprehensive collection of biological pathways (cf. **Section 4.2**). RegulatorTrail is a web-interface providing access to different methods to identify and prioritize key transcriptional regulators with respect to their impact on expression changes caused, for example, by pathological processes. NetworkTrail is a web-interface providing access to different analysis tools for regulatory networks (cf. **Section 4.4**).

While the previous three tools are of general purpose and can be well used for basic science and biomarker identification, the remaining two tools, DrugTargetInspector and ClinOmicsTrail^{bc}, focus on translational cancer research and clinical decision support for personalized medicine. DrugTargetInspector (DTI) is an interactive assistance tool for treatment stratification. DTI analyzes genomics, transcriptomics, and proteomics data sets and provides information on deregulated drug targets, deregulated biological pathways, and subnetworks, as well as mutations and their potential effects on putative drug targets and genes of interest (cf. **Chapter 5**). ClinOmicsTrail^{bc} is an interactive visual analytics tool for breast cancer treatment stratification. The web service offers rich functionality for the integration and analysis of clinical markers as well as transcriptomics and (epi-)genomics data sets with respect to a broad spectrum of biological, pharmacological, and medical knowledge. To this end, ClinOmicsTrail^{bc} provides a comprehensive assessment of a variety of treatment options based on the tumor's main driver mutations, the overall tumor mutational burden, activity patterns of core breast cancer-relevant pathways, drug-specific predictive biomarkers, the status of molecular drug targets, and pharmacogenomic implications (cf. **Chapter 6**).

Furthermore, in order to provide a comprehensive view on the data set(s) under investigation, our web services are seamlessly integrated with each other. Hence, once omics data sets are uploaded to one of the web services, they can also be readily analyzed using other applicable tools within the Graviton framework.

4.2 GeneTrail2 - a web service for multi-omics enrichment analysis

GeneTrail2 is the successor of the GeneTrail web service, which had been developed by Backes *et al.* [401] in 2007. The GeneTrail2 rewrite has mainly been conducted by Daniel Stöckel, Tim Kehl, and Patrick Trampert. I contributed by integrating additional resources and connections to our other web services to facilitate the use of GeneTrail2 in the context of personalized medicine. GeneTrail2 is published in *D. Stöckel, T. Kehl, P. Trampert, L. Schneider et al. Multi-omics enrichment analysis using the GeneTrail2 web service. Bioinformatics (2016) 31.10. doi: 10.1093/bioinformatics/btv770.*

The measurement and analysis of comprehensive multi-omics data sets allow studying the mechanisms of pathogenic processes in the initiation and progression of complex diseases such as cancer at unprecedented breadth and depth. While the analysis of such data sets has the potential to improve the diagnosis, prognosis, and therapy of diseases [402–405], the interpretation, validation, and translation of the obtained findings into clinical practice remains a big challenge [406]. The (epi-)genetic and cellular heterogeneity within tissues and across patients is one of the reasons why individual markers and even sets of marker genes oftentimes lack the robustness required for clinical applications [407]. Moreover, marker genes typically are selected independently from each other, neglecting their coordinated functioning within protein complexes and signaling cascades. As a remedy, methods for the pathway- and network-based analysis of omics data sets (cf. **Section 3.3.3**) have become a more and more popular tool for biomarker identification [326, 408–412]. In order to provide a wide variety of over-representation and pathway-enrichment methods to the research community, we developed GeneTrail2, a web service for the integrated analysis of genomics, transcriptomics, miRNomics, and proteomics data sets. In the following sections, we will first describe related tools and web services (**Section 4.2.1**). Afterward, we will describe the GeneTrail2 workflow and the offered functionality (**Section 4.2.2**). Finally, we will demonstrate the capabilities of GeneTrail2 in a case study for the identification of a treatment-relevant subtype in pancreatic cancer (**Section 4.2.3**).

4.2.1 Related work

There are numerous enrichment-based approaches for the functional annotation of gene sets and multi-omics measurements. Please refer to **Section 3.3.3** for an elaborate overview. Most of these approaches are complemented by corresponding standalone tools or web services, which will be briefly discussed in the following.

Several of those tools offer interfaces for specific enrichment methods, for example, DAVID [413] or GoMiner [414] for Over-Representation Analysis (cf. **Section 3.3.3.1**), whereas GSEA-P [415] and the Broad Institute [416] provide functionality for Gene Set Enrichment Analysis (cf. **Section 3.3.3.2**). There are also other tools available that offer more than one type of enrichment analysis, for example, Babelomics [417], GOrilla [418], or GOstat [419]. These tools, however, typically trade this additional functionality off against smaller numbers of available databases to test for [400]. Additional web services offering enrichment analyses are listed in the OMICtools database [420].

4.2.2 Workflow and functionality

As shown in **Chapter 3**, a large variety of enrichment methods exists, none of which can be considered as a 'magic bullet' applicable to all scenarios [421–423]. With GeneTrail2, we provide one of the most comprehensive collections of statistical methods and integrated *a priori* knowledge for enrichment analyses. GeneTrail2 is a complete rewrite of its predecessor, the GeneTrail [401] web service.

GeneTrail2 follows the modular framework approach for Gene Set Enrichment Analysis as presented by Ackermann and Strimmer [332], which consists of the following steps: (i) the identification of entity-level scores (e.g., of differential gene expression), (ii) the application of a set-level statistic to determine enrichment scores, and (iii) the significance assessment of the results. To this end, GeneTrail2 provides a comprehensive collection of biological categories to test for, numerous statistical tests for the identification of deregulated genes and pathways, and multiple views on the computed results. **Figure 4.3** provides an overview of the GeneTrail2 workflow.

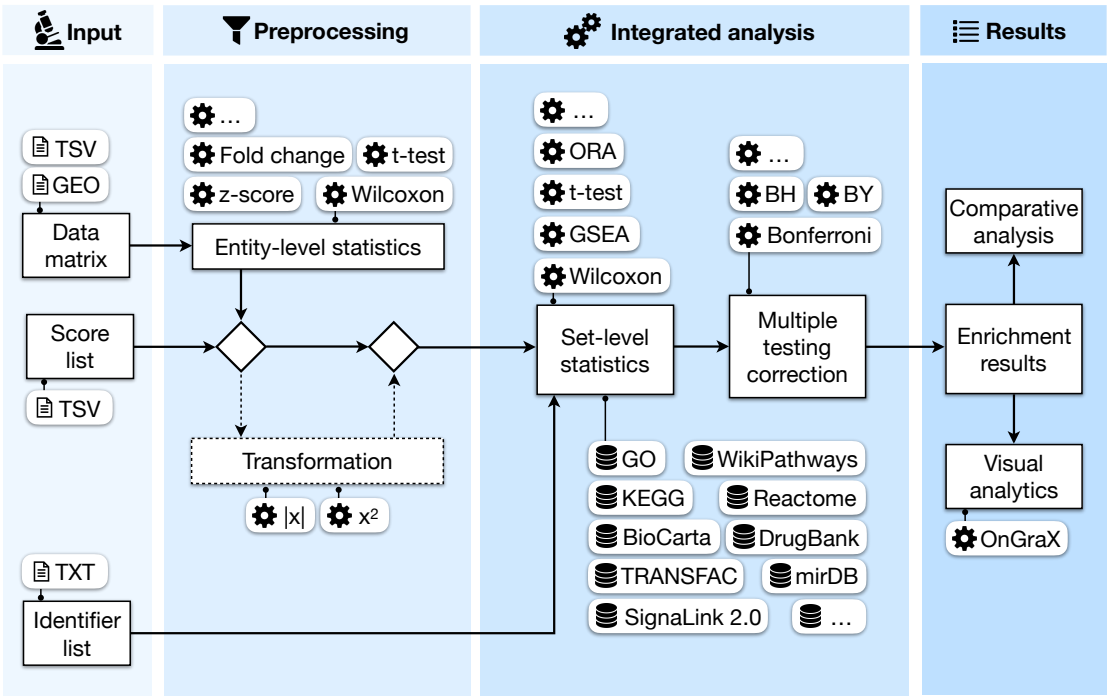


Figure 4.3 GeneTrail2 workflow. Input data formats are indicated by the 'file' icon, used databases by a 'database' icon, and statistical as well as computational methods by a gear wheel. Dashed arrows and boxes indicate optional steps. Due to space constraints, not all provided methods and databases are listed. Please refer to **Section A.4** for a complete list. BH: Benjamini-Hochberg adjustment, BY: Benjamini-Yekutieli adjustment, DTI: DrugTargetInspector, GEO: Gene Expression Omnibus, GO: Gene Ontology, GSEA: Gene Set Enrichment Analysis, HGNC: HUGO Gene Nomenclature Committee, KEGG: Kyoto Encyclopedia of Genes and Genomes, NT: NetworkTrail, ORA: Over-Representation Analysis, TSV: Tab-Separated Values, TXT: plain text file. The displayed icons were obtained from [15].

Data upload and scoring: First, the user uploads the data to be analyzed. GeneTrail2 supports genomics, transcriptomics, miRNomics, and proteomics data sets and can convert between 32 identifier types (cf. **Table A.3**). In cases where a data matrix is provided, the samples have to be

divided into a sample set and a reference set based on which entity-level scores are computed. In total, 13 identifier-level statistics are implemented (cf. **Table A.4**). After scoring, simple transformations such as taking the absolute value of the score or to square the result can be applied.

Enrichment analysis: In the next step, the user can select one of ten set-level statistics (cf. **Table A.5**) and the biological categories and pathways that should be analyzed. For human alone, GeneTrail2 provides more than 46,000 categories collected from over 30 databases including GO [424], KEGG [282], Reactome [425], WikiPathways [426], DrugBank [427], TRANSFAC [275], and miRDB [428] (cf. **Table A.7**). Moreover, custom user categories can be uploaded to GeneTrail2 in Gene Matrix Transposed (`GMT`) file format (cf. **Section A.4.2.1**). In order to account for multiple hypothesis testing (cf. **Section 3.3.1.1**), eight p-value adjustment methods are provided (cf. **Table A.6**). For each step of the analysis pipeline, the user can adjust the parameters of the employed methods. However, we also provide default values that should be applicable for most use cases.

Visualization of the results: The analysis results are provided in a ranked list of relevant pathways, ordered by confidence values (multiple testing-corrected p-values). The result list can be searched, sorted, and filtered. In order to make the results as meaningful and interpretable as possible, we provide various views for the enrichment results (cf. **Figure 4.4** and **Section A.4.3**). Besides the default view of a list of enriched or depleted pathways, GeneTrail2 also provides an *inverse enrichment* view. Here, differentially expressed genes are listed in decreasing order of their score of deregulation. For each gene, the pathways and gene sets the gene belongs to are listed and they can be investigated with respect to their enrichment status. For the integrative analysis of enrichment results from multiple omics data sets, GeneTrail2's *comparative enrichment* view can be used. This specialized view allows comparing several enrichment results side-by-side. Currently, there are two modes for comparison: intersection and union. The intersection mode only displays categories that are significantly enriched in all performed enrichment analyses, the union displays any category that is significantly enriched at least once. For the visual analytics-based investigation of dependencies between enriched or depleted categories, GeneTrail2 also offers a *dependency wheel* visualization and is integrated with the interactive graph visualization tool OnGraX [429]. The dependency wheel provides a circular representation of altered categories with connecting ribbons indicating the number of shared genes between two categories. OnGraX, on the other hand, provides a network visualization of significantly enriched or depleted categories, in which closely related categories form clusters.

Interoperability: Due to the fact that GeneTrail2 is based on the Graviton framework, it is also tightly integrated with our other web services. Specifically, once entity-level scores for multi-omics data sets are obtained, these scores can be forwarded to NetworkTrail (cf. **Section 4.4**) for network analysis or to DrugTargetInspector (cf. **Chapter 5**) for an assessment of deregulated drug targets and potential treatment options.

GeneTrail2 has been well received in the research community. On average, GeneTrail2 is used for more than 900 analysis runs each month. At the time of writing, it has been cited 69 times. In the following section, we will demonstrate GeneTrail2's capabilities in fostering translational research and biomarker identification by the example of pancreatic cancer.

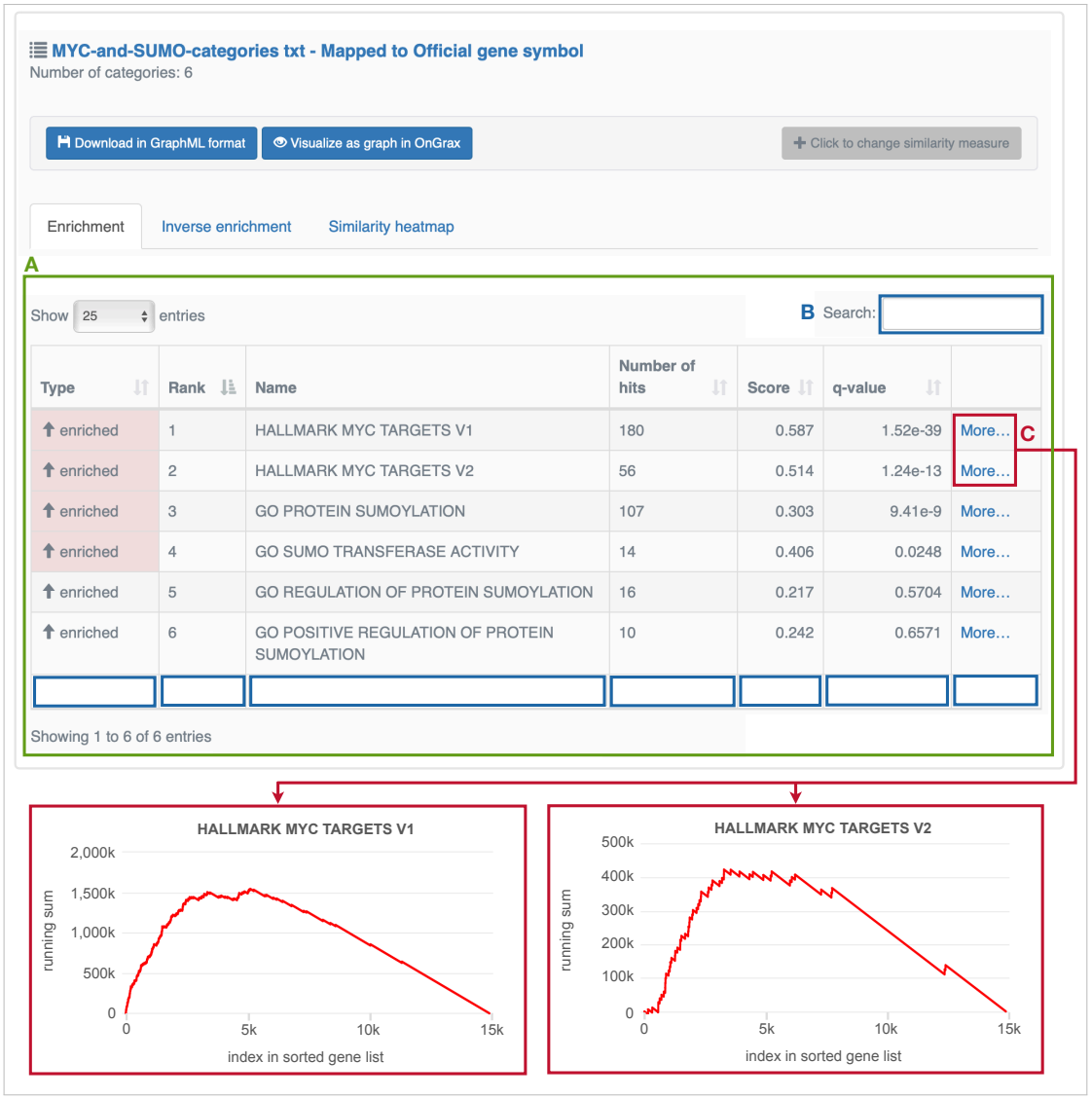


Figure 4.4 GeneTrail2 results for SUMO^{high} vs. SUMO^{low} subtype in PDAC. Enrichment results for the case study described in **Section 4.2.3**. Scores of differential gene expression between the two groups were computed using the independent shrinkage t-test. For the enrichment, unweighted GSEA with exact p-value computation was used. The obtained p-values were FDR-adjusted to a significance level of 0.05 using the Benjamini-Yekutieli method. **A) Main results table.** The tested categories are ranked by the significance score of their enrichment. Significantly enriched categories are highlighted in red. The table can be re-ordered by clicking on the respective 'sort symbol' next to the column name. **B) Searching and filtering.** The results table can be searched using the search box in the top right corner of the table. The results can moreover be filtered using the filter boxes below each of the table columns. **C) Running sum plots.** Clicking on the *More* button yields a visualization of the running sum of the corresponding category. Here, the running sums for the two most significantly enriched categories 'HALLMARK MYC TARGETS V1' and 'HALLMARK MYC TARGETS V2' are displayed.

4.2.3 Case study: The SUMO pathway as a therapeutic option in pancreatic cancer

Parts of this section are published in *Biederstädt, A., Hassan, Z., Schneewels, C., Schick, M., Schneider, L. et al. SUMO Pathway Inhibition Targets an Aggressive Pancreatic Cancer Subtype. Gut (2020) doi: 10.1136/gutjnl-2018-317856.* For the manuscript, I acquired data sets and performed numerous bioinformatics analyses, including the ones presented in this section. The presented drug sensitivity analyses were conducted by Alexander Biederstädt and colleagues.

Although pancreatic cancer only accounts for about 3% of all diagnosed cancers, it is with a 5-year survival rate of only 9% currently the fourth leading cause of cancer death in the United States [430]. With an incidence rate of 95%, Pancreatic Ductal AdenoCarcinoma (PDAC) is the most common form of pancreatic cancer [431]. Treatment options for PDAC are limited. Due to the lack of comprehensive preventive screening facilities, PDAC is typically only diagnosed in advanced stages. In combination with a relatively high median age of 71 years of PDAC patients at diagnosis, this leads to the fact that only 15-20% of PDACs can be resected [432]. Moreover, there are only few chemotherapeutic treatment options available for PDAC, none of which currently employs molecular biomarkers for treatment stratification [433].
With the goal of determining a potential novel stratified PDAC therapy, we analyzed gene expression data for a cohort of PDAC samples. Within this cohort, we were able to identify an aggressive molecular subtype that seems to be driven by a coactivation of MYC and the SUMO pathway. This dependence can potentially be therapeutically exploited for a subtype-specific treatment of pancreatic cancer. In the following sections, we will describe our analysis in more detail.

MYC and SUMOylation in PDAC: Several molecular subtypes of PDAC have been described in the literature, of which the so-called 'squamous', sometimes also called 'basal-like' or 'mesenchymal', subtype is associated with especially poor prognosis [434–436]. The squamous subtype is characterized by an activated MYC pathway [435]. The oncogenic transcription factor MYC is known to drive tumor initiation and progression in various cancer types [437, 438]. While there is also research focusing on directly targeting MYC for cancer treatment [439, 440], it has been shown that MYC exerts its oncogenic potential via the activation of growth-promoting downstream processes [441, 442]. Hence, tackling these effector processes might be a promising strategy for a stratified PDAC treatment, following the concept of 'synthetic lethality'. Synthetic lethality here describes the observation that the presence of a cancer-driving aberration (e.g., an MYC amplification) is accompanied by an increased vulnerability to perturbations of other molecular factors [443, 444]. The activation of the SUMOylation pathway has in this context already been described as a prerequisite for MYC-driven tumorigenesis [445] and several SUMOylation pathway genes have already been identified as synthetic lethal interactions for MYC (i.e., MYC-dependent tumor cells have been shown to be susceptible to suppression of the respective SUMO pathway genes) [446]. Moreover, MYC-driven SUMOylation has been shown to be a therapeutic vulnerability in B-cell lymphoma [447]. SUMOylation is a post-translational modification (cf. **Section 2.1**) in which members of the small ubiquitin-like modifier (SUMO)

protein family are conjugated to lysine residues of their target proteins. SUMOylation is involved in a variety of cellular processes, including the regulation of protein subcellular localization, the interaction of proteins, and DNA repair [448]. The covalent attachment of SUMO family members (SUMO1, SUMO2, SUMO3) to their target genes is mediated by a multistep catalytic process that involves various enzymes: SAE1 (SUMO1 activating enzyme subunit 1), UBA2 (ubiquitin-like modifier activating enzyme 2), UBE2I (ubiquitin-conjugating enzyme E2 I), PIAS1-4 (protein inhibitor of activated STAT 1-4), and the SENP protein family of SUMO-specific peptidases.

SUMOylation-based molecular subtype of PDAC: To further investigate the relevance of the SUMOylation pathway in PDAC, we analyzed a gene expression data set of a cohort of 96 PDAC patients provided by Bailey *et al.* [435]. The provided normalized gene expression scores were z-transformed (cf. **Section 3.3.2**) for each sample in comparison to all the other samples. For a predefined set of SUMO-related genes of interest, we investigated the scores of deregulation within the cohort and could identify a PDAC subtype ($SUMO^{high}$) with increased expression of core SUMO pathway genes, see **Figure 4.5**.

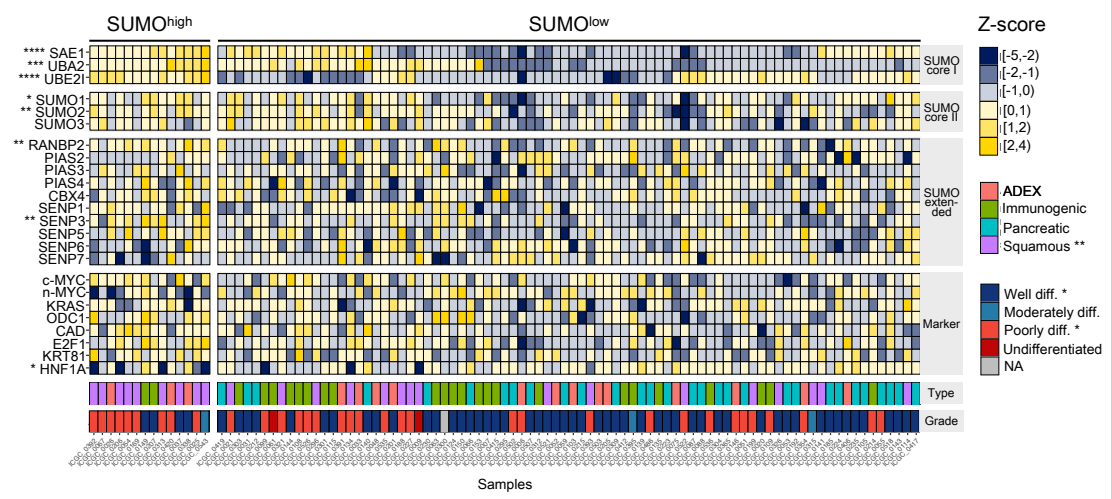


Figure 4.5 Differential gene expression defines $SUMO^{high}$ subtype in PDAC. Scores of differential intra-tumor gene expression are mapped to six colors between blue (downregulation) and yellow (upregulation). The $SUMO^{high}$ subtype is characterized by positive z-scores for SAE1, UBA2, and UBE2I (n=14). Differences in gene expression between the $SUMO^{high}$ and the $SUMO^{low}$ group were assessed using Student's *t*-test. The subtype and grade of the investigated samples are depicted in the last two 'rows' of the visualization. The enrichment of the respective clinical attributes in one of two groups was assessed using Fisher's exact test. The obtained significance levels are indicated using asterisks next to the gene name: $* < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001$.

With respect to clinical attributes, $SUMO^{high}$ PDACs are characterized by an enrichment of squamous subtypes (Fisher's exact test, p=0.0013) and poorly differentiated cells (Fisher's exact test, p=0.036).

In order to provide further evidence for the activating role of MYC in the $SUMO^{high}$ subtype, we performed enrichment analysis using GeneTrail2. To this end, we uploaded the gene expression matrix of PDAC samples provided by Bailey *et al.* and compared the $SUMO^{high}$ with the $SUMO^{low}$ group using the independent shrinkage *t*-test. Based on these scores of differential gene expression between the two groups, we tested for the enrichment of two sets of MYC-target

genes as provided by the Molecular Signatures DataBase (MSigDB) [449] and four categories related to SUMOylation obtained from the Gene Ontology [281]. To this end, we used the unweighted GSEA approach with exact p-value computation [327]. The obtained p-values were FDR-adjusted to a significance level of 0.05 using the Benjamini-Yekutieli method. The results, which are displayed in **Figure 4.4**, show a significant enrichment of MYC hallmark target genes, as well as the core SUMOylation gene sets in the $SUMO^{high}$ subtype.

Moreover, using Kaplan-Meier plots (cf. **Figure 4.6**), we determined that $SUMO^{high}$ PDACs are characterized by decreased progression-free as well as overall survival, and hence actually define an aggressive subtype of PDACs.

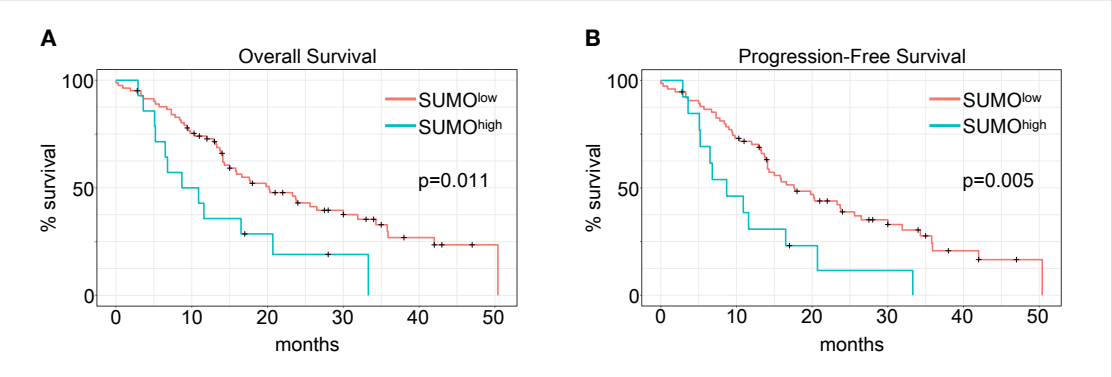


Figure 4.6 Kaplan-Meier plot for $SUMO^{high}$ vs. $SUMO^{low}$ subtype in PDAC. The follow-up status of the patients was divided into six types: 'alive - without disease', 'alive - with disease', 'alive - disease status unknown', 'deceased - of disease', 'deceased - of other cause', and 'deceased - of unknown cause'. Samples without any follow-up information were not considered. **A) Overall survival analysis.** Here, all samples are considered. **B) Progression-free survival analysis.** Here, only samples with the follow-up status 'alive - without disease' and all samples with follow-up status 'deceased' are considered.

Targeting the SUMO pathway in PDAC: In order to determine whether the SUMOylation machinery is indeed a relevant target in PDAC, we tested the activity of two small-molecule inhibitors of the SUMO-activating enzyme (SAE), a heteromer formed by the two subunits SAE1 and UBA2 (which sometimes also called SAE2). The inhibitors ML-792 and ML-93 selectively block SAE and hence prohibit SUMOylation. ML-792 has already previously been described to potently decrease cancer cell proliferation and work especially well in MYC hyperactive cells [450].

We analyzed the sensitivity of six human PDAC cell lines to the treatment with ML-792 and ML-93, respectively. Three of those cell lines showed low levels of MYC protein expression (BxPC-3, MIA-Pa-Ca-2, IMIM-PC1) and three had higher MYC levels (DAN-G, PaTu-8988T, PSN1). **Figure 4.7** provides an overview of the results. In comparison to ML-792, ML-93 induces a higher reduction in cell viability, which means that ML-93 is more effective than ML-792 for these PDAC cell lines (cf. **Figure 4.7 A**). Moreover, ML-93 shows an increased potency in the MYC-high PDAC cell lines, both with respect to the maximal response at a concentration of 1 μM (cf. **Figure 4.7 B**), as well as the reduction of colony size at a concentration of 40 nM (cf. **Figure 4.7 C**).

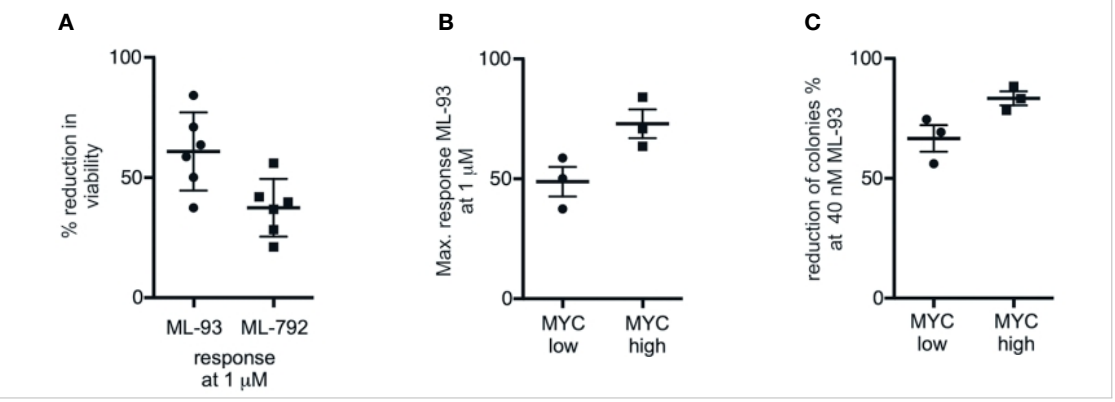


Figure 4.7 **Efficacy of SUMO inhibitors ML-792 and ML-93 in PDAC. A) Reduction of viability.** Boxplots for reductions in cell viability when treating PDAC cell lines with either ML-792 or ML-93 at 1 μM. **B) Response rates for ML-93.** Boxplots for response rates of six cell lines with low and high MYC protein expression levels, respectively, to treatment with 1 μM ML-93. **C) Colony reduction for ML-93.** Boxplots for colony reduction rates after treatment with 40 nM ML-93 in MYC-low and MYC-high cell lines, respectively.

To summarize, we provided evidence for the existence of a molecular subtype (SUMO^{high}) of PDAC that is defined by a co-activation of the SUMO pathway and the oncogene MYC. The SUMO^{high} subtype is characterized by an aggressive progression, poor prognosis, and the current lack of treatment options. To this end, we investigated two SUMO inhibitors as potential stratified treatment option. While the analysis of the relatively small number of cell lines can only serve as a first proof-of-concept, the results nevertheless indicate that MYC hyperactivation generates a vulnerability that potentially can be exploited by SUMO inhibitors.

4.3 RegulatorTrail - a web service for the identification of key transcriptional regulators

The RegulatorTrail web service is published in *T. Kehl, L. Schneider et al. RegulatorTrail: a web service for the identification of key transcriptional regulators. Nucleic Acids Research (2017) 45.W1. doi: 10.1093/nar/gkx350.* The web service was mainly been developed by Tim Kehl. I contributed via the implementation of the web service's analysis workflow for personalized medicine research scenarios. Moreover, I supported the conduction of the case studies as well as the writing and revision of the manuscript.

In the previous section, we presented GeneTrail2, a general-purpose tool for the identification of altered biological pathways and pathological processes. In order to obtain further mechanistic insights into complex diseases like cancer, the key regulatory elements that induce these pathological processes have to be identified. One essential class of regulatory elements are transcriptional regulators. Transcriptional regulators like transcription factors, coregulators, and epigenetic modifiers control the transcriptional machinery in eukaryotic cells and hence play major roles in most biological processes. Consequently, alterations in their structure, abundance, and activities have been associated with a variety of diseases, including cancer [451]. In this context, transcriptional regulators are commonly described as oncogenes or tumor

suppressors [452]. Prominent examples are the tumor suppressor gene TP53 and the oncogene MYC, which have been shown to be frequently altered in a variety of cancer types [437, 453]. Moreover, the capability of such regulatory elements to control the transcription of a large number of genes makes them interesting candidates to be targeted in cancer therapy [454, 455]. In order to identify those transcriptional regulators that are involved in pathogenic processes, we developed the web service RegulatorTrail. The tool provides eight different methods to identify and prioritize influential regulators on the basis of epigenomics and transcriptomics data. In the following sections, we will first briefly describe related approaches for the identification and prioritization of key transcriptional regulators (**Section 4.3.1**). Afterward, we will give an overview of RegulatorTrail's workflow and functionality (**Section 4.3.2**). As one of the methods provided by RegulatorTrail, we propose REGulator-Gene Association Enrichment (REGGAE), a novel approach to prioritize transcriptional regulators based on the combination of regulator-target interactions with enrichment analysis (**Section 4.3.3**). Finally, we will present a case study in which we use RegulatorTrail to assess the role of TCF3 as a potential master regulator in blastemal Wilms tumors (**Section 4.3.4**).

4.3.1 Related work

There are various approaches that aim at identifying and prioritizing those transcriptional regulators that might explain the differences in gene expression between two phenotypes (e.g., diseased vs. control). Most of the proposed approaches rely on *a priori* knowledge of transcription factors and their corresponding target genes as provided by various databases (cf. **Section 3.2.1**).

A first class of methods determines those regulators whose target genes show a significant enrichment in a set of differentially expressed genes using Over-Representation Analysis (cf. **Section 3.3.3.1**). To this end, TFactS [456] employs the hypergeometric test. The R-package DCGL [457] provides two different methods: (i) Targets' Enrichment Density (TED) tests for the enrichment of a regulator's targets in a list of deregulated genes using the binomial test and (ii) Targets' Differentially Co-Expressed Links Density (TDD) computes for a transcription factor T_i the 'density' of co-expressed target genes among all target genes:

$$TDD(T_i) = \frac{2k}{N \cdot (N-1)},$$

with N being the number of targets of T_i and k being the number of target genes that are differentially co-expressed with their regulator T_i, forming what Liu *et al.* call 'Differentially Co-Expressed Links' (DCLs) [458].

Another group of approaches is based on correlations between regulators and their target genes. The Regulatory Impact Factor metrics RIF1 and RIF2 investigate the co-expression between a regulator and its target genes [459]. The Correlation Set Analysis (CSA) [460] method aims at unveiling essential regulators by calculating the mean of all pair-wise correlations in the target set of a specific regulator. We recently developed an enrichment-based method called REGGAE that prioritizes regulators based on correlations within transcriptomics data, which will be presented in **Section 4.3.3**. Moreover, there are several other approaches that employ graph algorithms (e.g., TFRank [461]) or machine learning approaches (e.g., MIPRIP [462], Regulatory Snapshots [463]).

Besides the analysis and prioritization of transcriptional regulators on the basis of known, experimentally determined regulator-target interactions, another set of approaches focuses on the genome-wide prediction of transcription factor binding sites and motifs, which thereupon can be used to determine key regulators. Several methods have been proposed that jointly analyze epigenetic data (e.g., of open chromatin regions) and known transcription factor binding motifs. Examples for such approaches are CENTIPEDE [464], MILLIPEDE [465], or the more recently proposed method TEPIC [466]. The transcription factor binding affinities predicted by these tools can also be used to build models of gene expression that weight the transcriptional regulators by their relevance. An example of such an approach is the INVOKE (IdeNtification Of Key transcriptional regulators using Epigenetics data) analysis presented by Schmidt *et al.* [466, 467].

Many of the tools and methods described above are tailored to a specific application scenario. In order to provide an easy-to-use computational platform covering multiple research scenarios with respect to transcriptional regulation, we developed the web service RegulatorTrail, whose workflow and functionality will be described in the following section.

4.3.2 Workflow and functionality

RegulatorTrail provides eight different methods for the identification and prioritization of transcriptional regulators that cover the different methodological classes sketched in the previous section. To this end, RegulatorTrail offers solutions for four main use cases (i.e., 'scenarios'), which will be individually described in the following paragraphs. An overview of the different workflows is provided in **Figure 4.8**.

Due to the fact that the methods provided by RegulatorTrail require *a priori* biological knowledge of regulators and their target genes as well as regulator binding site motifs, we integrated data from several databases and resources. Formally, we define a pair consisting of a regulator and one of its experimentally determined target genes as Regulator-Target-Interaction (RTI). In order to provide a comprehensive list of known RTIs, we obtained data from seven databases: ChEA [468], ChIP-Atlas (`http://chip-atlas.org`), ChipBase [469], ENCODE [470], JASPAR [471], SignaLink [472], and TRANSFAC [275]. Information on regulator binding motifs is typically provided in the form of Position Count Matrices (PCMs). PCMs consist of four rows (one for each nucleotide) and one column for each position of the binding motif. The entries of a PCM are the number of occurrences of each nucleotide at the respective position in the motif [473]. For PCMs, we collected data from four databases: HOCOMOCO [474], JASPAR [471], the Kellis Lab ENCODE Motif database [475], and TRANSFAC [275].

Scenario I: In the first scenario, which is indicated by orange arrows and boxes in **Figure 4.8**, Over-Representation Analysis (ORA, cf. **Section 3.3.3.1**) is used to identify those transcriptional regulators whose target genes show a significant overlap with a gene set of interest. Here, users can either upload a set of interesting genes or provide a score file (cf. **Section 4.1.3**), which can be used to filter for, for example, the most differentially expressed genes. Moreover, users can choose a collection of RTIs from our database or upload a set of custom regulator-target interactions. To perform the ORA, three different statistical tests are offered: the hypergeometric test [456], Fisher's exact test [300], and the binomial test [457]. Finally, one of eight p-value

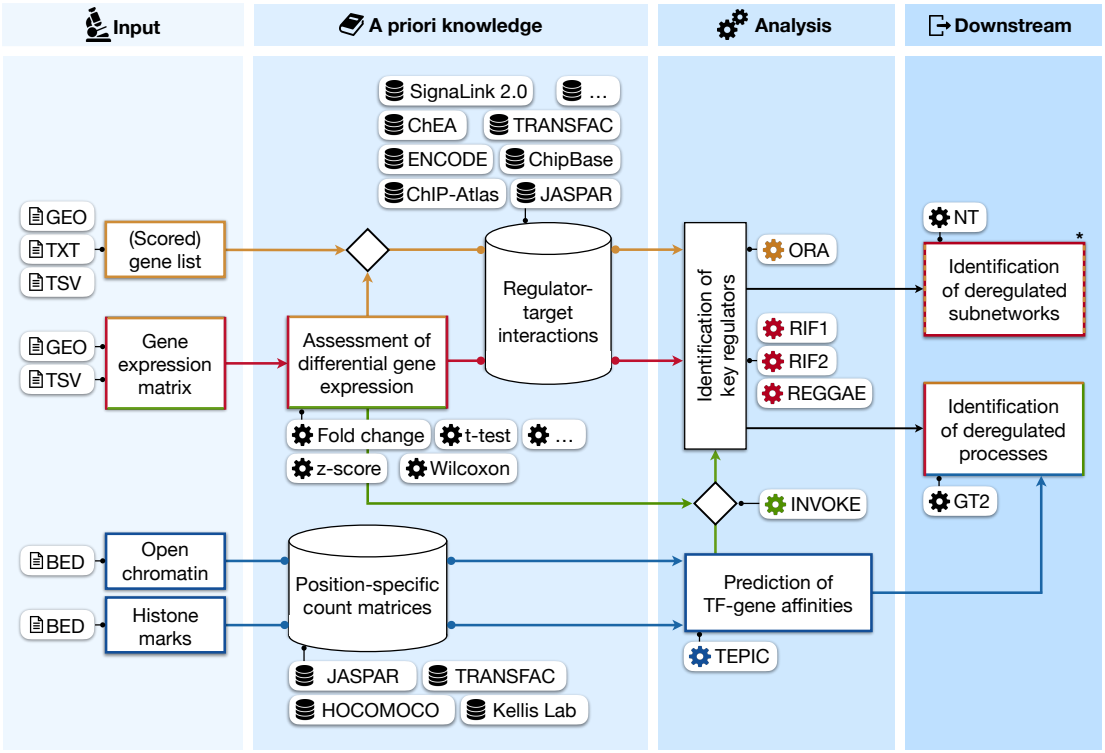


Figure 4.8 **RegulatorTrail workflow.** Input data formats are indicated by the 'file' icon, used databases by a 'database' icon, and statistical as well as computational methods by a gear wheel. The workflow, data types, and analyses used in the four types of research scenarios discussed in **Section 4.3.2** are color-coded: Orange: Scenario I, red: Scenario II, blue: Scenario III, and green: Scenario IV. The asterisk in the 'downstream' box indicates that a downstream analysis with NetworkTrail is only possible for Scenario I if a scored gene list was provided as input. BED: Browser-Extensible Data, GEO: Gene Expression Omnibus, GO: Gene Ontology, GT2: GeneTrail2, NT: NetworkTrail, ORA: Over-Representation Analysis, REGGAE: REGulator-Gene Association Enrichment, RIF1/2: Regulator Impact Factor method 1/2, TF: Transcription Factor, TSV: Tab-Separated Values, TXT: plain text file. The displayed icons were obtained from [15].

adjustment methods is applied (cf. **Table A.6**), resulting in a list of potentially influential regulators, sorted by their adjusted p-values.

Scenario II: In the second scenario, users can either upload a matrix of normalized gene expression values or use RegulatorTrail's integrated functionality to import data from GEO (cf. **Section 4.1.3**). Provided that the matrices contain samples belonging to two groups of interest (e.g., disease and control), differential gene expression between the two groups can be assessed. To this end, RegulatorTrail provides a variety of methods (cf. **Table A.4**). Based on the obtained deregulation scores, the user can select lists of up- or downregulated genes for further investigation (e.g., the top 250 most upregulated genes). For the identification of influential regulators, these genes can either be used to perform a target gene enrichment as described in **Scenario I**, or serve as input to the correlation-based methods RIF1, RIF2 [459], or REGGAE (cf. **Section 4.3.3**). The latter three methods have the advantage that they additionally provide information on whether the regulator has an activating or repressing effect on their targets. Also in this scenario, users can either use our predefined collection of RTIs or upload their own set of RTIs. The respective analysis steps are highlighted in red in **Figure 4.8**.

Scenario III: The third scenario (indicated in blue in **Figure 4.8**) aims at predicting transcription factor binding sites via the use of known binding motifs. In order to restrict the search space, open chromatin data are used as input. To this end, candidate regions of open chromatin in the form of DNase-hypersensitive sites [476] or H3K4me3 histone modification data [477] can be uploaded in `BED` format (cf. **Section A.1.5**). In order to extract those genomic regions from the open chromatin data that overlap with transcriptional start sites of genes, RegulatorTrail considers a window of user-defined size for each gene. These windows are centered at the respective gene's most 5′ Transcriptional Start Site (TSS). The thereby identified candidate regions are then analyzed using the segmentation-based method TEPIC [466], which uses a customizable collection of Position Count Matrices (PCMs) to predict transcription factor-gene affinities. The resulting matrix of TF-gene affinity scores can also be used to build a predictive model of gene expression (cf. **Scenario IV**).

Scenario IV: As already indicated in the previous paragraph, the transcription factor-gene affinity scores computed by a TEPIC analysis can also be used to predict those regulators that have the highest impact on the expression of their target genes. The corresponding analysis steps are highlighted in green in **Figure 4.8**. In addition to a `BED` file containing open chromatin regions, which also serves as input in **Scenario III**, a score file containing gene expression measurements for the same sample is required. The TF-gene affinity matrix $A \in \mathbb{R}_0^{+\,n \times m}$ consisting of affinity scores for all pairs of n genes and m transcription factors is used in a linear regression model to predict the gene expression of the n genes $g = (g_1, g_2, ..., g_n)$:

$$g = A \cdot \beta + \epsilon,$$

with the regression coefficients $\beta = \{\beta_1, \beta_2, ..., \beta_m\}$ and the error term $\epsilon = \{\epsilon_i, \epsilon_2, ..., \epsilon_n\}$. In order to control the regression coefficients, the user can choose between three types of regularization: Ridge [478], Lasso [479], and the Elastic Net [480], which each applies a different penalty to the regression coefficients β_i. The linear models yield a list of features with non-zero

regression coefficients, which are likely to play essential roles in the transcriptional regulation of the analyzed sample. Moreover, model performance is assessed by Pearson correlation [343], Spearman correlation [481], and the mean-squared error [482] between the predicted and the actual gene expression.

The results from all four scenarios either yield lists of interesting regulators (**Scenarios I**, **II**, and **IV**) or putative target genes (**Scenario III**), which can further be analyzed using GeneTrail2 (cf. **Section 4.2**) to assess the potentially common functional context of the identified regulators and targets, respectively. Moreover, in cases where scores of differential gene expression were either uploaded to or computed within RegulatorTrail (**Scenarios I** and **II**), these scores can also seamlessly be used for the identification of deregulated regulatory subnetworks via NetworkTrail (cf. **Section 4.4**).

4.3.3 REGGAE - REGgulator-Gene Association Enrichment

The work described in this section is published in *Kehl, T., Schneider, L. et al. REGGAE: a novel approach for the identification of key transcriptional regulators. Bioinformatics (2018) 1.8. doi: 10.1093/bioinformatics/bty372.* The REGGAE method was developed by Tim Kehl and Hans-Peter Lenhof. I contributed to the presented case studies and the writing and revision of the manuscript.

For the identification and prioritization of transcriptional regulators that have a strong influence on the expression of a given set of genes, we have proposed an alternative approach to the ones described in **Section 4.3.1**: REGulator-Gene Association Enrichment (REGGAE) analysis combines association scores between regulators and their target genes with a Gene Set Enrichment approach to identify and prioritize the influence of the investigated regulators on expression changes between two phenotypes.
In the following paragraphs, we will first describe the methodology of REGGAE in more detail and then compare REGGAE's capability to identify relevant regulators with the ones of competing tools. Moreover, in **Section 4.3.4**, we will present a case study in which REGGAE was used to investigate transcriptional regulators in an aggressive subtype of Wilms tumor.

The REGGAE algorithm: **Figure 4.9** provides an overview of the REGGAE workflow and **Algorithm 4.1** shows the REGGAE algorithm in pseudocode.
The input for a REGGAE analysis consists of a normalized gene expression matrix $\mathbf{E} \in \mathbb{R}^{p \times n}$ of measurements for p genes in n samples, where the n samples belong to two phenotypes (e.g., disease and control). In a first step, scores of differential gene expression between these two groups are computed using one of the numerous entity-level statistics offered by REGGAE (cf. **Table A.4**). Based on these scores, the genes are sorted and either the most up- or down-regulated genes (based on a user-defined threshold) are considered for further analysis. The separate consideration of up- and down-regulated genes is required as transcriptional regulators could affect some of their target genes in an activating manner, while others are repressed. If not considered individually, these effects may cancel each other out.

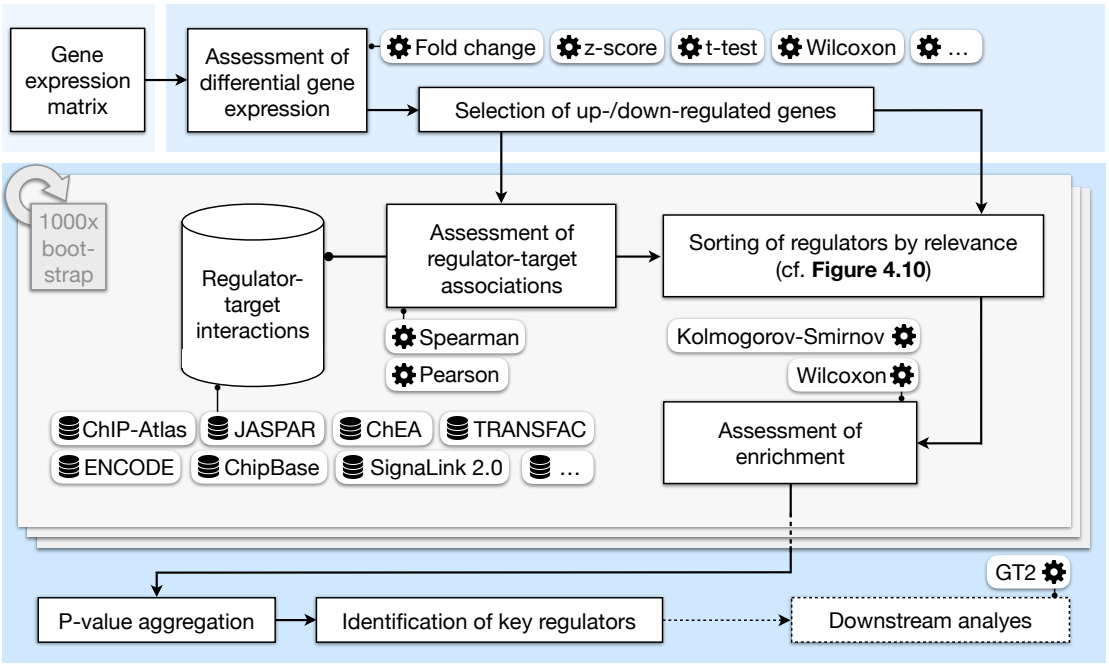


Figure 4.9 REGGAE workflow. The blue background boxes correspond to the stages of the method, ranging from data input over scoring and preprocessing to the actual analysis and options for downstream analysis. The used databases are indicated by a 'database' icon and statistical as well as computational methods by a gear wheel. GT2: GeneTrail2. The displayed icons were obtained from [15].

For better readability, we consider in the following only one of the two (up- or downregulated) sorted gene lists: $D = \{g_1, g_2, ..., g_m\}$. Based on a collection of regulator-target interactions (cf. **Section 4.3.2**), those l_i regulators $R_{g_i} = \{r_{i1}, r_{i2}, ..., r_{il_i}\}$ that can influence the expression of a specific gene $g_i \in D$ are considered. For every pair of regulator r_{ij} ($j \in \{1, ..., l_i\}$) and target gene g_i, we calculate the correlation between the expression values of regulator and target across all samples using either Pearson's correlation coefficient [483] for linear dependencies or Spearman's rank correlation coefficient [481] for non-linear dependencies. For each gene g_i, the regulator list R_{g_i} is sorted with respect to the (absolute or signed) correlation coefficients, which is considered as the degree of regulator-target association (cf. 'rows' in **Figure 4.10 A**).

Based on the sorted list of differentially expressed genes $D = \{g_1, g_2, ..., g_m\}$ (first 'column' in **Figure 4.10 A**) and their corresponding sorted regulator lists $R^*_{g_i} = \{r^*_{i1}, r^*_{i2}, ..., r^*_{il_i}\}$, we create a new list $L = \{r^*_{11}, r^*_{21}, ..., r^*_{m1}, r^*_{12}, r^*_{22}, ...\}$. This new list L is created by traversing the sorted list of differentially expressed genes D in decreasing order. First, the most strongly associated regulators r^*_{i1} for each gene g_i are added to L, followed by the second most strongly associated regulators r^*_{i2} for each gene g_i and so on (cf. **Figure 4.10 B**).

In the final list L, regulators are sorted by their impact on their target genes, this means that regulators that are strongly associated with highly deregulated genes will occur at the beginning of the list. Hence, regulators with a major impact on the observed differential gene expression should be enriched at the top of the list. In order to assess and quantify such an accumulation, we carry out an enrichment analysis for each regulator individually, as indicated in **Figure 4.10** by the blue regulator r_1. To this end, REGGAE offers the Wilcoxon rank-sum test [319] or the unweighted version of the Kolmogorov-Smirnov test [328]. The resulting p-values are adjusted

using the method proposed by Benjamini and Yekutieli [311] (cf. **Section 3.3.1.1**). Finally, REGGAE provides a list of regulators sorted by their adjusted p-values.

Input: Normalized gene expression matrix $\mathbf{E} \in \mathbb{R}^{p \times n}$ for p genes and in total n disease and control samples, lists of regulators R_{g_i} for each gene g_i, overall number of regulators s

Output: List M^* of regulators sorted by their adjusted p-values

$D_{all} \leftarrow$ computeDifferentialGeneExpressionBetweenTumorAndControl($\mathbf{E}$)
$D \leftarrow$ considerOnlyDeregulationInOneDirectionAndSort(D_{all}, 'up')
foreach *gene* g_i *in* D **do**
  $C_{g_i} \leftarrow [\,]$ //List of correlations of g_i with its regulators
  foreach *regulator* r_{ij} *in* R_{gi} **do**
    $C_{g_i} \leftarrow$ append(C_{g_i},computeCorrelation($\mathbf{E}_{g_i}$,$\mathbf{E}_{r_{ij}}$))
  end
  $R^*_{g_i} \leftarrow$ sortRegulatorsInDecreasingOrderOfCorrelation(R_{g_i}, C_{g_i})
  //This yieds $R^*_{g_i} = \{r^*_{i1}, r^*_{i2}, ..., r^*_{il_i}\}$
end
$L \leftarrow [\,]$ // List of regulators to be used for enrichment
for j *in* 1 *to* s **do**
  for i *in* 1 *to* $|D|$ **do**
    $L \leftarrow$ append(L,r^*_{ij})
  end
end
$M \leftarrow [\,]$ //List of enrichment results for all regulators
for j *in* 1 *to* s **do**
  $M_j \leftarrow$ performEnrichmentOnEachRegulator(L, j)
end
$M^* \leftarrow$ adjustForMultipleTestingAndSort(M)

Algorithm 4.1 **Pseudocode for REGGAE algorithm.** For exemplary purposes, we consider only upregulated genes in this pseudocode, hence the parameter 'up' in the function 'considerOnlyDeregulationInOneDirectionAndSort'. The variables $\mathbf{E}_{g_i}$ and $\mathbf{E}_{r_{ij}}$ stand for the measured gene expression values for a gene g_i and a regulator r_{ij}, respectively. For the functions 'computeDifferentialGeneExpressionBetweenTumorAndControl', 'computeCorrelation', 'performEnrichmentOnEachRegulator', and 'adjustForMultipleTestingAndSort', users can select from several options, see description in main text.

However, due to the fact that technical noise in the gene expression measurements can bias the computed correlations, which are an essential part of the REGGAE algorithm, we offer the use of a following bootstrapping scheme [484] to improve the robustness of the method. This means that we perform numerous runs of the REGGAE algorithm, each time on a slightly altered input data set and average the results. To this end, we create b bootstrap samples, where each sample is generated by randomly selecting (with replacement) n columns from the original gene expression matrix $\mathbf{E}$. By this, we obtain an input matrix of the original dimensions, but with moderately varying content. Next, REGGAE is applied to each of the b bootstrap samples. Finally, the median p-value of the b results is used as the final score for this regulator.

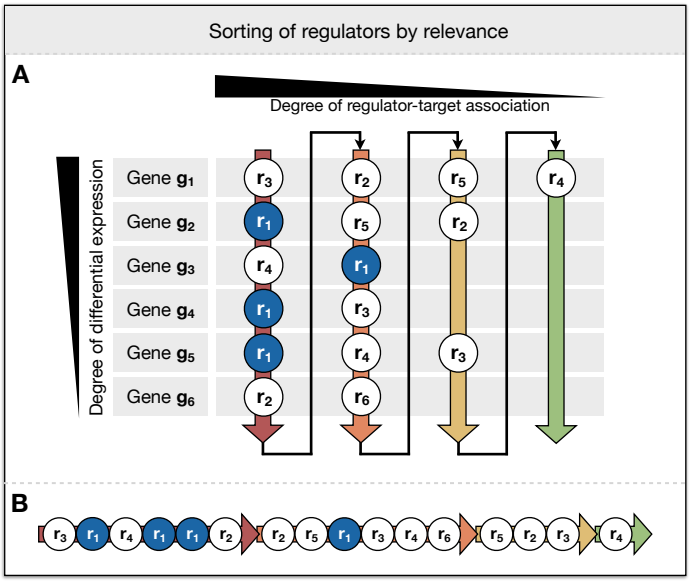


Figure 4.10 **Sorting of regulators by relevance**. **A)** A subset of either up- or downregulated genes $\{g_1, g_2, ..., g_m\}$ are sorted according to their degree of differential expression. For each gene g_i, the list of targeting regulators is sorted with respect to the degree of regulator-target association. The blue nodes correspond to a regulator under investigation. **B)** Shows the sorted list of regulators that is obtained by concatenating the regulators in a column-wise manner. The obtained list serves as input for enrichment analyses with the respective regulators as 'category' under investigation. Figure adapted from [485].

Comparison to other methods: In order to compare the capabilities of REGGAE with competing methods, we applied REGGAE and seven other methods (CSA, RIF1, RIF2, TDD, TED, TFactS, and TFRank, see **Section 4.3.1**) to a breast cancer data set and investigated whether we could identify key regulatory factors involved in breast cancer initiation and progression.

Breast cancer is one of the most common types of cancer and the second leading cause of cancer death among women [486]. One of the clinically most relevant breast cancer subtypes are estrogen receptor-positive (ER+) tumors, which comprise around 70% of diagnosed cases [487] and generally have a better prognosis than estrogen receptor-negative (ER-) tumors [488] (see also **Section 6.2**).

We applied REGGAE and the other methods to a data set of 37 breast cancer cell lines, which was published by Heiser *et al.* [489] and for which we obtained their estrogen receptor status from a study by Neve *et al.* [490] (cf. **Section A.5.1.1**).

In total, we compared 16 ER+ and 21 ER- cell lines to find those transcriptional regulators that have a strong influence on the differential expression between the two phenotypes. As a first step, we assessed differential gene expression between ER+ and ER- samples using the shrinkage *t*-test [318]. The genes were sorted with respect to their t-scores. From the sorted list, we selected several gene sets for further investigation: all genes that were significantly up-regulated in ER+ tumors (with $p < 0.01$), as well as the top 250, 500, 750, and 1,000 genes. We individually applied REGGAE to the five lists and obtained five ranked lists of regulators. These five rankings were aggregated into the final result using sum-of-ranks. For each regulator, we used the maximum of the five p-values from the five runs as overall p-value. Parameters for all analyses and corresponding results can be found in **Section A.5.1.2**. Runtimes for all methods are depicted in **Table A.9**.

Table 4.1 shows the top five regulators identified by REGGAE (columns 1 and 2) and the results for these regulators as obtained from the other methods (columns 3-9). The first entry in each table cell either contains p-values or scores if no p-values are provided by the respective method (indicated by an asterisk). The second entry (in parentheses) indicates the rank of the gene in the respective result lists.

Regulator	REGGAE	CSA	RIF1*	RIF2*	TDD	TED	TFactS	TFRank*
FOXA1	$6.34 \cdot 10^{-141}$ (1)	$9.76 \cdot 10^{-6}$ (359)	−2.87 (116)	8.34 (18)	$8.4 \cdot 10^{-6}$ (956)	1.0 (843)	1.0 (953)	6.92 (2)
GATA3	$3.23 \cdot 10^{-137}$ (2)	$9.76 \cdot 10^{-6}$ (421)	−2.73 (113)	5.16 (62)	$8.7 \cdot 10^{-6}$ (747)	1.0 (681)	0.05 (369)	6.56 (3)
ESR1	$6.52 \cdot 10^{-129}$ (3)	$9.76 \cdot 10^{-6}$ (509)	−1.93 (229)	−0.1 (915)	$8.4 \cdot 10^{-6}$ (949)	1.0 (440)	1.0 (790)	10.28 (1)
MYB	$6.34 \cdot 10^{-125}$ (4)	$9.76 \cdot 10^{-6}$ (262)	−2.07 (130)	4.14 (75)	$8.4 \cdot 10^{-6}$ (878)	1.0 (606)	0.31 (519)	5.45 (6)
SPDEF	$2.6 \cdot 10^{-118}$ (5)	$9.76 \cdot 10^{-6}$ (40)	−3.05 (32)	8.54 (15)	$1.4 \cdot 10^{-5}$ (434)	1.0 (892)	$3.6 \cdot 10^{-19}$ (72)	6.44 (4)

Table 4.1 **Top five regulators in ER+ breast cancer identified by REGGAE in comparison to other approaches**. *For REGGAE, CSA, and TFactS adjusted p-values are depicted. *For RIF1, RIF2, and TFRank, which do not provide p-values, the respective test statistic value is shown. Numbers in parentheses represent the rank in the sorted result list.*

The top five regulators identified by REGGAE are FOXA1, GATA3, ESR1, MYB, and SPDEF, all of which have already been described as prognostic markers in breast cancer indicating a good prognosis [491–494]. Especially, FOXA1, GATA3, and ESR1 have been reported as co-located and co-expressed in breast cancer cells [495, 496]. Moreover, FOXA1, GATA3, ESR1, and SPDEF are reported as master regulators in fibroblast growth factor (FGF) signaling and breast cancer risk in ER+ cells [497]. The top five REGGAE candidates have also been identified by CSA and TFRank as significant. Notably, with respect to the rankings of the top candidates, REGGAE and TFRank yield very similar results that differ strongly from the remaining methods. TFactS detected only two of the five regulators as significant, RIF1 and RIF2 detected four out of the five among their top 200 results.

To summarize, our results indicate that most methods identify similar key regulators, however, with substantially different rankings. Although most methods were able to assign at least some of the central regulators of ER+ cells as being relevant, REGGAE and TFRank excelled in terms of the actual ranking of those regulators.

Besides this breast cancer case study, we also analyzed perturbation gene expression signatures of induced MYC overexpression in mouse lymphomas and knock-out experiments in human embryonic stem cells, in which REGGAE outperformed TFRank. Please refer to the corresponding publication for additional details [485].

While the results described above demonstrate that REGGAE is able to yield reasonable results, we also used REGGAE to obtain novel biological insights into the regulatory mechanisms underlying an aggressive subtype of Wilms tumors. The corresponding analysis and results will be presented in the following section.

4.3.4 Case study: The role of TCF3 as potential master regulator in blastemal Wilms tumors

The results described in this section were published in *T. Kehl, L. Schneider et al. The role of TCF3 as potential master regulator in blastemal Wilms tumors. International Journal of Cancer (2019) 144.6. doi: 10.1002/ijc.31834.* The performed analyses were conceptualized and conducted by Hans-Peter Lenhof and Tim Kehl. I contributed to the interpretation of the results as well as the writing and revision of the manuscript.

4.3.4.1 Wilms tumors

Wilms tumors (WT), also known as nephroblastomas, are pediatric malignant tumors and the predominant type of childhood kidney cancer [498]. Although Wilms tumors generally have a good prognosis with survival rates over 90%, some subtypes are associated with a high risk of relapse [499]. Typically, WTs mainly consist of three histological components: blastema, stroma, and epithelial cells. The proportions and degree of differentiation of these cell types can strongly vary between tumors [500].

For the treatment of WTs, two different schemes have been established. While the Children's Oncology Group (COG) does not see a need for routine preoperative treatment, children treated according to the protocol of Société Internationale d'Oncologie Pediatrique (SIOP) typically undergo neoadjuvant chemotherapy. Preoperative chemotherapy can strongly affect the composition of cell types in the (remaining) tumor. In this context, a larger amount of surviving, chemoresistant blastema (inducing the so-called 'blastemal subtype') confers a high risk [501]. In order to better understand the role of blastema as a high-risk factor in WTs, it is of utmost importance to elucidate the regulatory mechanisms that differentiate blastemal from non-blastemal components of WTs.

4.3.4.2 REGGAE analysis

In order to identify transcriptional regulators that potentially explain the differences between blastemal and non-blastemal Wilms tumors, we collected and analyzed 33 WT samples from patients that were treated according to the SIOP protocol, which means that they received neoadjuvant chemotherapy with actinomycin-D, vincristine and, in the case of metastases, doxorubicin.

The data set contains biopsies of 17 blastemal and 16 non-blastemal tumors (cf. **Table A.12**). Gene expression was assessed using Agilent SurePrint arrays. The corresponding data set has been published on the Gene Expression Omnibus platform (accession number: GSE98334).

In order to identify the most influential regulators, we first assessed differential gene expression between the blastemal and non-blastemal tumor samples using shrinkage *t*-test (cf. **Section 3.3.2**). The genes are then sorted in descending order of their t-scores.

For the analysis using REGGAE, we created ten different lists: Based on a significance threshold of 0.01, we selected all significantly upregulated genes (538) and all significantly downregulated genes (317). Moreover, we created eight lists containing the 250, 500, 750, and 1,000 most

upregulated and most downregulated genes, respectively. In a next step, we applied REGGAE to each of those lists. The results for the five lists of upregulated genes and analogously for the five lists of downregulated genes were aggregated using the second-order statistic for p-values [319]. The aggregated p-values are finally FDR-adjusted using the approach by Benjamini and Yekutieli [311].

Table 4.2 gives an overview of the ten most significant regulators based on the lists of upregulated and downregulated genes, respectively. Please refer to **Section A.5.2** for the full list of identified regulators.

Upregulated genes		Downregulated genes	
Regulator	P-value	Regulator	P-value
RUNX1 (−)	$1.22 \cdot 10^{-180}$	NR2F2 (−)	$7.83 \cdot 10^{-116}$
TCF3 (+)	$5.96 \cdot 10^{-163}$	MAX (+)	$3.27 \cdot 10^{-105}$
NR2F2 (+)	$6.19 \cdot 10^{-163}$	TCF3 (−)	$3.12 \cdot 10^{-95}$
MAX (−)	$3.54 \cdot 10^{-157}$	RUNX1 (+)	$1.78 \cdot 10^{-94}$
SFPQ (+)	$1.06 \cdot 10^{-136}$	CREBBP (−)	$8.51 \cdot 10^{-78}$
ELF1 (−)	$4.60 \cdot 10^{-134}$	ELF1 (+)	$1.09 \cdot 10^{-76}$
KDM5B (+)	$1.68 \cdot 10^{-131}$	SUMO2 (−)	$4.03 \cdot 10^{-74}$
HDAC1 (+)	$9.85 \cdot 10^{-125}$	CREB1 (−)	$4.42 \cdot 10^{-70}$
SIN3A (+)	$2.90 \cdot 10^{-123}$	SMC3 (−)	$8.33 \cdot 10^{-70}$
CREB1 (+)	$5.84 \cdot 10^{-123}$	UBTF (−)	$9.24 \cdot 10^{-61}$

Table 4.2 **Top ten regulators identified by REGGAE analysis in blastemal Wilms tumors.** Aggregated REGGAE results for upregulated and downregulated genes, respectively. Each ranking was obtained via a sum-of-rank aggregation of the REGGAE results for input lists of the following sizes: 250, 500, 750, and 1,000, as well as all significantly upregulated (538) and downregulated (317) genes (with p-value < 0.01). The colors of the gene symbols in the first and third column indicate whether the mean correlation coefficient between a regulator and its target genes is **positive** (+) or **negative** (−).

According to REGGAE, the most influential regulators for both up- and downregulated genes are NR2F2, TCF3, RUNX1, and MAX. The nuclear receptor subfamily 2 group F member 2 (NR2F2) is a transcription factor that is involved in the differentiation of human embryonic stem cells [502]. Moreover, it has been shown to play a role in tumor initiation and progression of several cancer types [503]. Transcription factor 3 (TCF3) is - as the generic name suggests - a transcription factor that plays essential roles in a variety of processes: It has been shown to induce gene expression of Wnt-responsive genes [504]. The Wnt signaling pathway is known to be activated in blastemal WTs [505] and it has been linked to tumorigenesis as well as chemoresistance in various tumor types [506, 507]. Along with TCF3, we have also identified two of its coactivators: CREBBP and EP300 (see extended results in **Table A.10**). The RUNX family transcription factor 1 (RUNX1) is a known tumor suppressor in breast cancer and acute lymphoblastic leukemia [508, 509]. It is involved in the differentiation of hematopoietic stem cells to lymphoid or myeloid cells [510]. The MYC associated factor X (MAX) is a transcription factor that forms different kinds of homo- and heterodimers with other transcriptional regulators like MYC, MNT, or MXI1 and which is involved in regulation of cell proliferation, differentiation, and apoptosis [511].

These results indicate that many of the essential regulators active in blastemal subtype WTs are involved in the regulation of stem cells and hence are likely to induce the stem-like character and potentially the aggressiveness of blastemal WTs. To validate this hypothesis, we investigated

the chromatin signaling network of mouse Embryonic Stem Cells (ESCs) [512]. The network contains many of the top regulators identified by REGGAE, including TCF3. The 49 genes belonging to this network are highly enriched for both REGGAE results lists. Moreover, all of those genes have binding sites for TCF3 and the majority of those genes also shows a strong absolute correlation ($|\rho| > 0.5$) in gene expression with TCF3. In order to confirm the role of TCF3 in blastemal WTs, we additionally verified that TCF3 target genes are significantly enriched in the set of most highly expressed genes in blastemal WTs.

To further support our observation of the stem cell-like character of blastemal WTs, we performed a comparison of histone marks in embryonic stem cells and Wilms tumor cells. Here, our results indicate that blastemal WT cells share several characteristics with ESCs that are not present in non-blastemal tumor cells. Additionally, we observed that TCF3 targets are again significantly enriched in the set of genes with activating histone marks in their promoter regions. This result reinforces our assumption that TCF3 is a crucial regulatory element in blastemal WTs. Please refer to the manuscript for details on the performed analyses and additional results.

To summarize, our results emphasize that stem cell-like properties are a central characteristic of blastemal Wilms tumors and might even foster the increased malignancy and chemoresistance of this tumor subtype. Specifically, our results highlight the role of TCF3 as a central element in a circuitry of regulatory and epigenetic mechanisms. Along with TCF3, we identified several additional biomarkers that are characteristic of the blastemal subtype. These insights can potentially be utilized to improve diagnosis, prognosis, and even therapy of patients with Wilms tumors.

4.4 NetworkTrail - a web service for identifying and visualizing deregulated subnetworks

The results described in this section were published in *D. Stöckel et al. NetworkTrail - a web service for identifying and visualizing deregulated subnetworks. Bioinformatics (2013) 29.13. doi: 10.1093/bioinformatics/btt204*. I was not involved in the initial development of this web service, yet I have contributed to the web service's maintenance ever since. Due to the fact that NetworkTrail is also integrated with the other tools described in this and the following chapters, we will briefly describe the functionality, but refer the reader for additional details and case studies to the above-mentioned manuscript.

The deregulation of signaling pathways plays a central role in many complex diseases, and especially in cancer (cf. **Chapter 2**). For the identification and elucidation of altered biological processes that characterize a given phenotype, numerous computational methods have been proposed (cf. **Section 3.3.3**). Some of these methods focus on the analysis of gene sets (cf. **Sections 3.3.3.1** and **3.3.3.2**), while others also take the topology of the underlying biological signaling network into account (cf. **Section 3.3.3.3**).

While GeneTrail2 (cf. **Section 4.2**) provides an easy-to-use web service for the gene set based analysis of biological categories and pathways, NetworkTrail focuses on the topology-based analysis of altered pathologic processes. The NetworkTrail web service enables users to detect the most deregulated pathways and subgraphs in biological networks.

4.4.1 Workflow and functionality

An overview of the analysis workflow within NetworkTrail is shown in **Figure 4.11**.

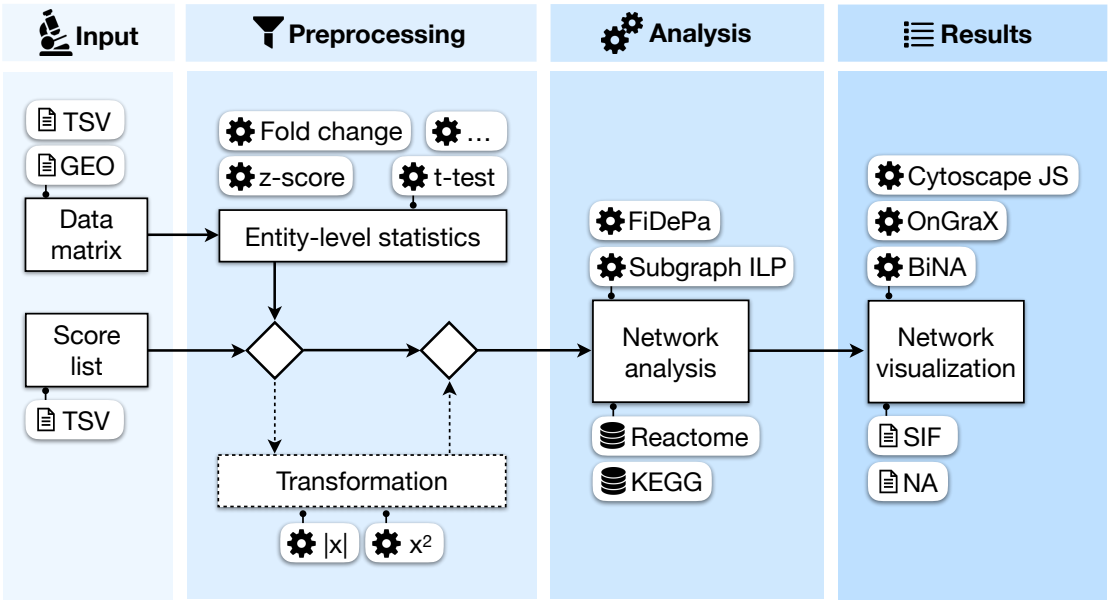


Figure 4.11 **NetworkTrail workflow**. Input (and output) data formats are indicated by the 'file' icon, used databases by a 'database' icon, and statistical as well as computational methods by a gear wheel. BiNA: Biological Network Analyzer, GEO: Gene Expression Omnibus, KEGG: Kyoto Encyclopedia of Genes and Genomes, NA: Node Attribute file, SIF: Simple Interaction File, TSV: Tab-Separated Values. The displayed icons were obtained from [15].

Similar to the web services presented in the previous sections, the analyses conducted by NetworkTrail are based on differential gene expression. Users can either provide scores of differential gene expression in the form of a score file or provide a gene expression matrix (cf. **Section 4.1.3**). In the latter case, users can choose from a variety of entity-level statistics to compute deregulation scores per gene (cf. **Table A.4**). These scores can optionally be transformed using basic operations like taking the absolute value.

In a next step, the scores of deregulation per gene are mapped onto the nodes of a signaling network derived from KEGG [282] (cf. **Section 3.2.2**). For the computation of the most deregulated subnetwork, NetworkTrail provides two algorithms: an Integer Linear Programming (ILP) formulation proposed by Backes *et al.* [362] and the FiDePa algorithm devised by Keller *et al.* [355]. Details on both methods can be found in **Section 3.3.3.3**.

The computed deregulated subgraphs can be visualized in three ways: (i) directly in the browser using a visualization based on Cytoscape.js [513], (ii) in a new browser tab via the OnGraX graph visualization tool [429], or (iii) locally on the user's computer via a Java web start provided by the Biological Network Analyzer (BiNA) [514]. Finally, a representation of the resulting subgraph in the form of `SIF` (Simple Interaction Format) and `NA` (Node Attribute) files (cf. **Section A.6.1**) can be downloaded for offline usage and visualization in graph visualization tools like the standalone version of Cytoscape [515].

A case study using the NetworkTrail functionality and additional details on the corresponding visualization using BiNA will be presented in **Chapter 5**.

5

DrugTargetInspector

Main parts of this chapter are published in *L. Schneider, D. Stöckel, T. Kehl et al. DrugTargetInspector: An assistance tool for patient treatment stratification. International Journal of Cancer (2016) 138.7. doi: 10.1002/ijc.29897*. The DrugTargetInspector web service was predominantly developed by myself. It is based on the Graviton software architecture devised by Daniel Stöckel and Tim Kehl.

Many complex diseases, and especially cancer, are caused by genetic and molecular aberrations that emerge in an evolutionary manner and manifest in various ways in the molecular, cellular, and ultimately phenotypic characteristics of the disease [516]. Due to the genetic and molecular heterogeneity of tumors, and the fact that cancerous clonal evolution can rapidly induce drug resistance, the treatment of cancer is still a grand challenge. The intrinsically positive fact that the number of chemotherapeutic agents is steadily growing (currently there are more than 200 FDA-approved anticancer drugs), however, renders the search for an optimal treatment even more difficult, in particular, if a combination therapy is required. Hence, in order to determine an optimal treatment for a given tumor, an in-depth characterization of the tumor's genetic and phenotypic makeup can provide a sound basis for decision-making.

5.1 Related work

In order to support systems medicine and translational research, several bioinformatics methods and tools have emerged over the last years that approach the pathological and pharmacological dependencies in complex diseases from various angles: The Drug-Gene Interaction database (DGIdb) [517] combines information from several databases like DrugBank [427], PharmGKB [518], or CancerCommons [519] to identify those drugs that target genes in a user-provided gene set. Similarly, the Search Tool for InTeractions of CHemicals (STITCH) [520] provides information on drug-target interactions, which are rated by a confidence score that is based on the occurrence of the respective interaction in other databases, the results of in-vitro experiments, and the literature. The web service canSAR [521] combines biological, pharmacological, and chemical data with biological network topologies to facilitate hypothesis generation for drug development. Recently, DrugTargetProfiler has been presented [522], a visual analytics tool for the interactive analysis and exploration of drug-target interaction networks obtained from the authors' own open data crowdsourcing portal for the annotation of molecules as drug targets [523]. However, all of these tools lack the integration of tumor-specific genomics or transcriptomics data.

The continuous development of high-throughput experimental techniques, which allow for the molecular characterization of diseases at an increasing resolution (cf. **Section 3.1**), has enabled the development of a variety of tools that utilize molecular data to elucidate treatment options. The ConnectivityMap [524] and its successor, the Library of Integrated Network-based Cellular Signatures (LINCS) [525], provide collections of gene expression profiles from numerous human cell lines, which were treated with a variety of perturbing agents, including more than 1,000 distinct bioactive small molecules. User-provided query signatures (gene expression profiles) can be uploaded and compared to all reference expression profiles using Gene Set Enrichment Analysis (GSEA) (cf. **Section 3.3.3.2**). By this, those perturbagens that are most strongly correlated or anti-correlated with the query can be identified. The ConnectivityMap is, amongst others, used by the tool DrugPairSeeker [526] to predict optimal pairs of drugs that potentially 're-regulate' cancerous gene expression profiles towards gene expression patterns of healthy cells.

Another class of tools tries to improve treatment selection via the investigation of deregulated pathway and network structures: The Cytoscape plugin OCSANA [527] aims at selecting an optimal and minimal combination of interventions that disrupt all regulatory and signaling pathways that exist between two gene sets of interest. These sets could, for example, be a set of genes with genomic aberrations and a set of genes that show differential expression. Other tools seek to simulate the behavior of cancer cells and hence their reaction to therapeutic agents. Iadevaia *et al.* [528] aim at identifying optimal drug combinations by tracing signaling cascades using phosphoproteomics data. The Oncosimulator [529] combines clinical and molecular data to simulate a tumor's response to treatment with a drug, including toxicologically relevant side effects. Other approaches predict the response of cancer cells to drug treatment by population modeling of Darwinian evolution [530, 531].

A major factor explaining the differences in drug responses between tumors of the same type, and even subtype, are the genomic alterations that drive the disease. Accordingly, genomic alterations are also used by several tools to predict treatment outcome. A comprehensive resource for the elucidation of pharmacogenomic interactions (i.e., mutations that affect the efficacy of a drug) is the Genomics of Drug Sensitivity in Cancer database (GDSC1000) [532, 533]. Similar information with a focus on drug targets is provided by the Cancer Drug Resistance database (CancerDR) [534], which contains pharmacological profiles of 148 anticancer drugs across 952 cell lines, including known mutations and their effect on treatment response. In 2014, the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project presented a challenge to predict the drug sensitivity of various breast cancer cell lines [299]. The participant's approaches ranged from principal component analysis over regression trees to ensemble methods. The best-performing team used a Bayesian multitask multiple kernel learning method [535]. Besides the mere prediction of drug sensitivity, there are also numerous approaches that aim at generating mechanistic insights into the molecular processes and dependencies that inform drug sensitivity. For example, Aben *et al.* devised TANDEM, a two-stage elastic net regression, which combines genomics and transcriptomics data to identify molecular key players predictive for drug response [536]. Similarly, in LOBICO, logic models of combinations of molecular aberrations are derived from the GDSC1000 data set [537].

In order to support systems medicine and translational research, we have developed DrugTargetInspector (DTI), an interactive assistance tool that provides rich functionality for the integrative analysis of tumor-specific omics data sets. In order to reveal the characteristics of a given tumor, DTI analyzes and integrates genomics, transcriptomics, and proteomics data sets, where genomics data sets provide information on genetic alterations (cf. **Section 5.2.2**). These alterations can affect the eligibility of therapy options in different ways. On the one side, driver mutations induce the deregulation of certain processes and pathways and hence the knowledge of such mutations and the induced pathway activities is important for decisions on (targeted) therapies. On the other side, mutations can make therapies ineffective or reduce their efficacy. To account for these dependencies, mutation data can be uploaded to DTI and, based on the GDSC1000 database, DTI annotates given mutations with their pharmacogenomic effects across a large panel of drugs (cf. **Section 5.2.3.2**). Transcriptomics and proteomics data sets can be used to identify deregulated signaling pathways and processes. To this end, DTI performs enrichment analyses on a large set of pathways derived from a variety of databases (cf. **Section 5.2.3.4**). Based on these pathway activities and the corresponding information on mutations, putative target pathways, drug targets and their corresponding drugs can be identified. DTI also determines if these drug targets are deregulated and offers functionality to determine their effect on downstream processes. To this end, DTI performs a subgraph analysis, which reveals the most deregulated subnetwork rooted in a drug target of interest (cf. **Section 5.2.3.4**). The subnetwork is visualized along with its corresponding gene expression data, which allows for a visual assessment of how the downstream molecules might be influenced by the root node. DrugTargetInspector is also fully integrated with its sister projects GeneTrail2 (cf. **Section 4.2**), RegulatorTrail (cf. **Section 4.3**), and NetworkTrail (cf. **Section 4.4**). In summary, this provides a powerful integrated tool suite for cancer therapy stratification by providing in-depth analyses of tumor omics data sets and a characterization of various aspects of dysregulation in the tumor, making DTI a valuable addition to existing clinical decision-support systems. DTI can be freely accessed at `https://dti.bioinf.uni-sb.de`.

In the subsequent sections, we will first give an overview of DTI's workflow and functionality (**Section 5.2**), followed by three case studies highlighting DTI's potential to foster treatment decision-making and translational research (**Section 5.3**).

5.2 Workflow and functionality

The identification of tumor-specific characteristics that can serve as a sound basis for treatment stratification requires (i) the integration of a broad range of heterogeneous omics data sets and databases, (ii) the development of powerful statistical methods for the analysis of high-dimensional and noisy data, and (iii) the generation of explorative tools that provide intuitive visualizations of relevant results. For the identification of tumor-specific characteristics relevant for an optimal treatment stratification, DTI integrates genomics and molecular data with *a priori* biological, pharmacological, and medical knowledge. Based on these data, several complementary analyses are performed, each of which yields a different view on the tumor under investigation. **Figure 5.1** provides an overview of DTI's workflow.

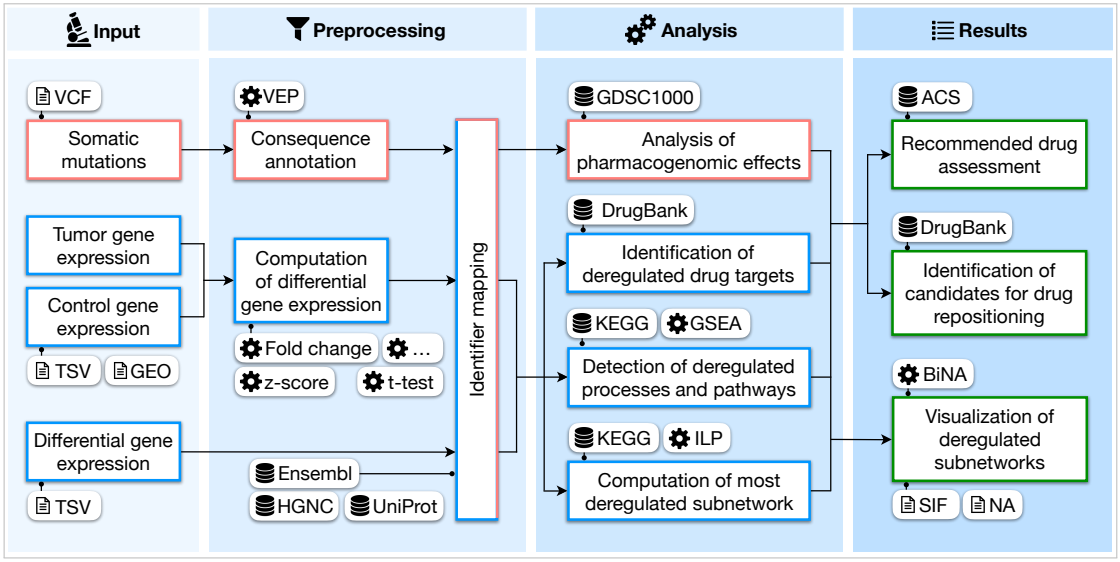


Figure 5.1 **DrugTargetInspector workflow**. The box border colors correspond to the type of data used in the corresponding step or analysis. Red: mutation data, blue: gene expression data. Please note that the term 'gene expression' here covers mRNA expression, miRNA expression, as well as protein levels. The green border color in the last column stands for results and potential analysis endpoints. Input (or output) data formats are indicated by the 'file' icon, used databases by a 'database' icon, and statistical as well as computational methods by a gear wheel. ACS: American Cancer Society, BiNA: Biological Network Analyzer, GDSC1000: Genomics of Drug Sensitivity in Cancer, GEO: Gene Expression Omnibus, GSEA: Gene Set Enrichment Analysis, ILP: Integer Linear Programming, KEGG: Kyoto Encyclopedia of Genes and Genomes, NA: Node Attribute file, SIF: Simple Interaction File, TSV: Tab-Separated Values, VCF: Variant Call Format, VEP: Variant Effect Predictor. The icons in this figure were obtained from [15].

DrugTargetInspector (DTI) incorporates various databases, including gene ontologies, regulatory network databases, specialized databases on miRNAs and transcription factors, pharmacological databases, and clinical decision guidelines, which are described in **Section 5.2.1**. In **Section 5.2.2**, the different data types and corresponding file formats that can be uploaded to DTI are specified. The variety of analysis tools provided by DTI is sketched in **Section 5.2.3**, followed by a description of the visualization of the analysis results in **Section 5.2.4**.

5.2.1 Integrated databases

DrugTargetInspector incorporates various databases, ranging from gene ontologies and regulatory network data over pharmacological data to clinical decision guidelines. For information on genes and the functional categories they belong to, we employ NCBI Gene [270] and GO [424]. The Kyoto Encyclopedia of Genes and Genomes (KEGG, cf. **Section 3.2.2**) [282] provides information on regulatory signaling cascades and their topologies. In order to map miRNAs and transcription factors to the genes they affect, we incorporate miRTarBase [538] and TRANSFAC [539]. Pharmacological data are obtained from DrugBank [427], which provides comprehensive information on therapeutic agents and their respective drug targets (cf. **Section 3.2.3**). Not only genomic variations in a drug's molecular target, but also mutations unrelated to the actual target can affect a drug's efficacy (e.g., activating mutations downstream of a receptor that is targeted by an inhibitor). To this end, DTI employs the Genomics of Drug Sensitivity in Cancer database (GDSC1000) [533] and PharmGKB [540], which contain information on pharmacogenomic interactions for a wide range of drugs. DTI also provides information about standard-of-care treatments and treatment decision guidelines for cancer. There are several guidelines available in the United States [541] and in Europe [542, 543] that are substantially overlapping. A list of 74 cancer subtypes and their respective recommended treatments was obtained from the American Cancer Society (ACS) [541]. For 32 of these subtypes, information about targeted treatment options was provided.

5.2.2 Tumor-specific input data

DrugTargetInspector enables the upload of omics data in several file formats. Transcriptomics data (mRNA/miRNA) and proteomics data can be uploaded as tab-delimited score files (`TSV`). Additionally, we support the upload of (gene) expression matrices. Expression data sets from the Gene Expression Omnibus (GEO, cf. **Section 3.2.3**) [544] can be imported by providing identifiers for GEO SEries (`GSE`) files or GEO Data Set (`GDS`) files. For the analysis of genetic alterations, mutation data can be uploaded in Variant Call Format (`VCF`, cf. **Section A.1.4**), which is a common format to describe variations identified in next-generation sequencing experiments (cf. **Section 3.1.2**) and which is supported by many tools for variant identification like SAMtools [545], GATK [546], or VarScan2 [547] (cf. **Section 3.1.2.3**). DTI supports numerous identifiers describing biological entities in the uploaded files: EntrezGene [270], HGNC symbols and IDs [272], KEGG [282], and UniProt identifiers [274] for genes and mirBase identifiers [278] for miRNAs.

5.2.3 Integrated analyses

Based on gene expression data of a tumor sample and one or several healthy controls, DTI offers various methods to calculate scores of differential expression per gene. This functionality is presented in **Section 5.2.3.1**. Alternatively, users can directly provide score files that describe the degree of deregulation of all genes (or microRNAs), which are used as basis for the analysis of deregulated drug targets, processes, and pathways in the following. In order to provide a detailed view of the genetic variations in the tumor, DTI analyzes the mutations contained in uploaded `VCF` files and assesses their pharmacogenomic effects (**Section 5.2.3.2**). Moreover, DTI analyzes deregulated drug targets (**Section 5.2.3.3**), provides a general overview on deregulated processes in the tumor (**Section 5.2.3.4**), and investigates the potential impact of targeted drugs on deregulated pathways and regulatory networks (**Section 5.2.3.5**).

5.2.3.1 Scoring and preprocessing

For the integration of the tumor-specific input data provided by users with the database content of DTI, identifiers types have to be harmonized. To this end, DTI uses Graviton's elaborate mapping functionality (cf. **Section 4.1.3**). As the internal representation of gene- and protein-based entities, HUGO gene symbols are used, and for miRNAs, miRBase identifiers are employed.

In order to identify characteristic deregulated processes and molecular key players of a tumor under investigation, DTI offers functionality for the calculation of scores measuring the degree of deregulation of all genes in the tumor tissue in comparison to one or several healthy samples (e.g., obtained from the healthy tissue surrounding the tumor). After the upload of a normalized (gene) expression matrix (or import of a `GSE` file from the GEO database), users can select and assign samples to a test set and a reference set. For therapy stratification, the test set would usually consist of an individual tumor sample and the reference set should contain healthy control sample(s). In this case, users can use (log) mean fold quotients [391] as scoring method to assess differential gene expression. In the case that the reference set consists of several samples, users can alternatively choose the z-score [392] as scoring method. If required, also larger sample groups can be compared against each other using one of the following tests: independent shrinkage *t*-test [393], independent Student's *t*-test [394], Wilcoxon-Mann-Whitney test [395], signal-to-noise ratio [396], *F*-test [397], and (log) mean fold quotient [391].

5.2.3.2 Analysis of pharmacogenomic effects

(Epi-)genetic variations, especially driver mutations, play a central role in tumor initiation and progression. In order to take these effects into account for the treatment decision-making process, somatic variant data can be analyzed in DTI. To this end, drug targets potentially affected by variations and drugs for which pharmacogenomic effects have been described by GDSC1000 are highlighted in the results. In order to also account for aberrations that have not (yet) been described as pharmacogenomic, DTI also analyzes the impact of the contained variations on the protein sequence (e.g., stop gained, missense, frameshift). To this end, Ensembl's Variant Effect Predictor (VEP) [548] is employed (cf. **Section A.2**). In terms of treatment stratification, genetic variations like mutations and SNPs are also of particular interest,

as they can have a major effect on the tumor's sensitivity to certain drugs [8]. To this end, drug targets potentially affected by variations and drugs for which pharmacogenomic effects have been described by GDSC1000 are highlighted in the results.

5.2.3.3 Identification of deregulated drug targets

At the time of writing, there are more than 200 FDA-approved anticancer drugs available, which are characterized in databases like DrugBank [427]. Many of those drugs have specific molecular targets and can be employed for targeted therapies [71, 549]. DTI offers functionality for identifying those genes that are (i) deregulated in the tumor under investigation, (ii) whose deregulation has a strong effect on the tumor's phenotype, and (iii) that can be targeted by known drugs. In a first step, based on the provided (or computed) scores of deregulation and drug-target interactions contained in DrugBank, DTI identifies all known drug targets that are significantly deregulated in the sample under investigation. In a second step, DTI analyzes the biological pathways containing these drug targets (**Section 5.2.3.4**) and molecular subnetworks potentially affected by them (**Section 5.2.3.5**).

5.2.3.4 Detection of deregulated processes and pathways

In order to investigate pathways affected by the considered drug target, DTI performs (unweighted) Gene Set Enrichment Analyses (cf. **Section 3.3.3.2**) for all KEGG pathways containing the corresponding gene. The analysis is based on the scores of deregulation for the genes contained in each pathway and reveals which regulatory pathways are most affected by the disease and potentially also by the deregulation of the drug targets. Significantly enriched (or depleted) pathways are highlighted in the results and can also be visually inspected on the KEGG website.
Moreover, in order to generally assess deregulated processes in the tumor, without the focus on a specific drug target, additional Gene Set Enrichment Analyses can be natively performed using GeneTrail2, which provides a variety of different enrichment algorithms and biological categories to test for (cf. **Section 4.2**).

5.2.3.5 Computation of most-deregulated subnetwork

In order to provide an even more in-depth view on the role of deregulated drug targets under investigation, DTI also assesses the potential downstream effects of deregulated drug targets. To this end, the KEGG regulatory signaling network is considered, whose nodes correspond to genes/proteins and whose edges describe regulatory interactions between these genes/proteins (cf. **Section 3.2.2**). The absolute values of the scores of differential expression, which were computed by DTI in a previous step, are mapped onto the network nodes. Based on this weighted network and a drug target under investigation, we calculate the connected subnetwork of size k that (i) is rooted in the considered drug target and (ii) maximizes the sum over all weights of the nodes, which corresponds to the connected subnetwork of size k with the highest degree of deregulation (cf. **Table A.11**). Besides the most deregulated subnetwork, DTI also provides the option to compute the most upregulated subgraph. In this case, the original scores of differential expression are used instead of the transformed ones.

We formulate this optimization problem as an extension of the Integer Linear Programming formulation (ILP) proposed by Backes *et al.* [362], which is presented in detail in **Section 3.3.3.3**. The ILP is solved using the Branch&Cut framework of CPLEX [550]. The degree of deregulation (or upregulation) of these subnetworks is assumed to mirror the phenotypic effect of the deregulated drug target and thus can help to judge the influence of its deregulation. In order to assess the robustness of the subgraph analysis results, not a single subgraph is computed, but a set of subgraphs with different sizes within a predefined range, which are used to generate a consensus graph (see also **Section 5.2.4.3**). Moreover, in addition to analyses on the downstream effects potentially induced by the drug target, DTI can also compute the most deregulated (or upregulated) subgraph that is upstream of a drug target of interest. By this, aberrant genes or mutations that might have induced the deregulation of the drug target itself can be elucidated.

5.2.4 Results visualization

After users have been guided through the data upload and scoring steps, they are forwarded to DrugTargetInspector's results page, which provides a prioritized list of deregulated drug targets and which serves as a starting point for additional in-depth analyses. In the subsequent sections, we will first describe DTI's results page in general (**Section 5.2.4.1**), followed by details on enrichment results (**Section 5.2.4.2**), and the visualization of deregulated subgraphs and genomic aberrations contained in the data set (**Section 5.2.4.3**).

5.2.4.1 Results page

Figure 5.2 shows DrugTargetInspector's results page for an exemplary colon cancer data set, which will be presented in detail in **Section 5.3.2**. The results page consists of two main elements: the table of deregulated drug targets and a side panel, in which filtering options and computation parameters can be set and from which additional analyses can be triggered. The results table is fully searchable and sortable and provides a prioritized list of molecular drug targets, decreasingly sorted by their score of differential expression. For each drug target (**Figure 5.2 A**), the second column indicates whether or not the drug target is affected by a mutation (**Figure 5.2 B**). Clicking on the blue 'map' icon will open a pop-up with additional details on the contained mutation(s) and their potential effects on the target. The third column contains the drug targets' scores of differential expression, followed by a list of drugs that target this gene/protein. A literature search in PubMed [551] for any combination of a drug and the corresponding drug target can be performed by clicking on the 'chevron' icon in the *Drugs* column (**Figure 5.2 C**). Next to the 'chevron' icon, there is a 'link' icon, which indicates whether or not pharmacogenomic interactions are known for the respective drug (**Figure 5.2 D**). Additional detailed information on the displayed molecular targets and drugs is available via links to respective database entries in NCBI Gene [270] and DrugBank [427]. Based on the subtype of cancer that is under investigation, standard-of-care drugs, as obtained from the American Cancer Society [541], are highlighted in green (**Figure 5.2 E**). The last column contains a 'magnifying glass' icon, which links to a menu, from which enrichments (cf. **Section 5.2.3.4**) and subgraph analyses (cf. **Section 5.2.3.5**) can be performed (**Figure 5.2 F**). Finally, the side panel (**Figure 5.2 G**) contains numerous collapsible boxes. Clicking on a box unfolds additional details and options, as well as parameter settings for the analyses. For

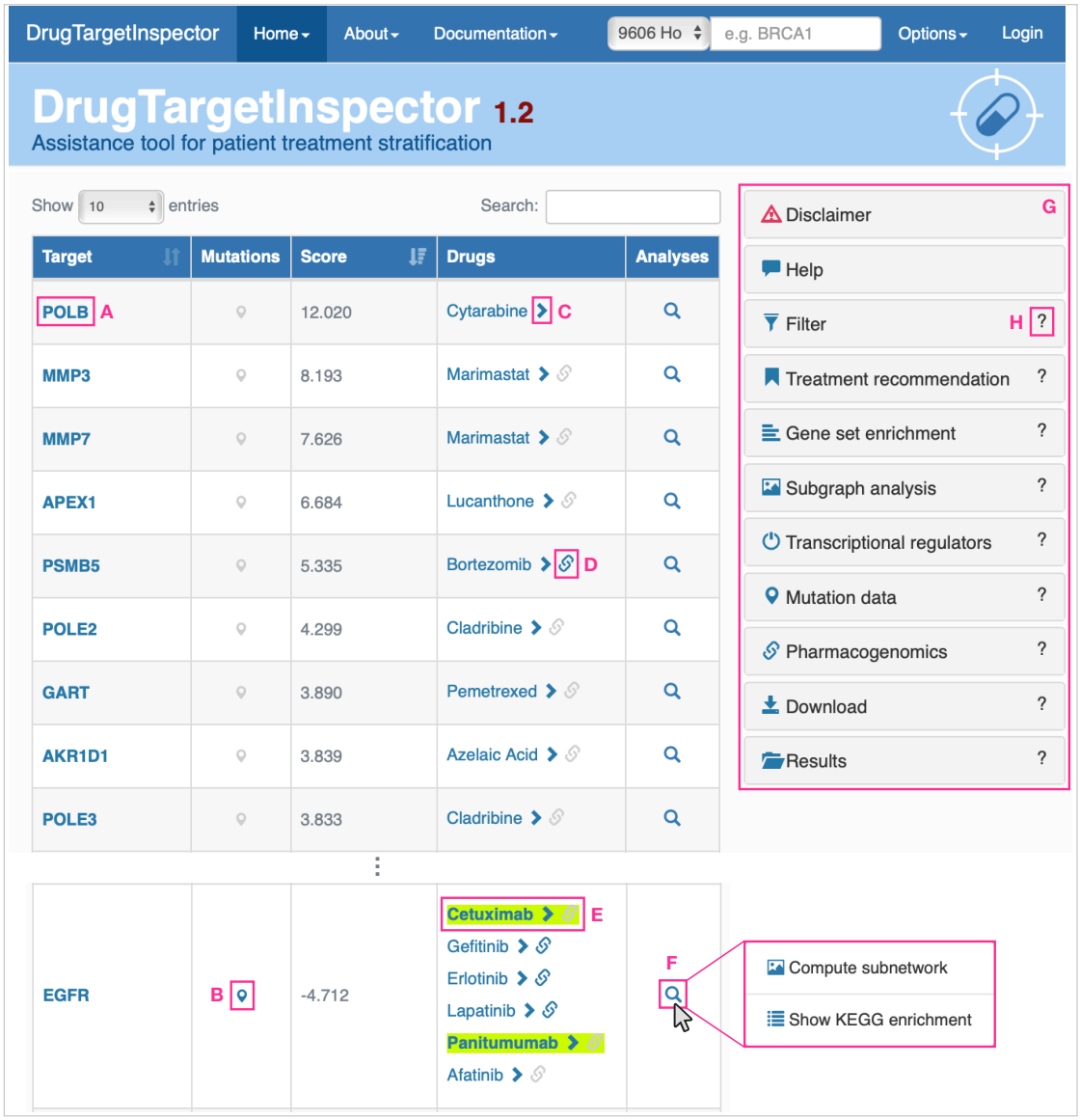


Figure 5.2 DrugTargetInspector results page. The main panel contains the table of deregulated drug targets, their mutation status, scores of differential expression, and a list of drugs targeting the respective molecules. **A) Drug target.** Details on the molecular drug target can be obtained by clicking on the target's name, which links to the respective entry in NCBI Gene. **B) Mutation indicator.** The 'marker' icon indicates whether (blue) or not (gray) the sample contains a mutation that affects the target gene. In cases where a mutation is present, clicking on the blue icon opens a window containing additional information. **C) Literature search.** Clicking on the 'chevron' icon links to the results of a joint literature search for the drug and the molecular target in PubMed. **D) Pharmacogenomic dependencies.** The 'link' icon indicates whether (blue) or not (gray) pharmacogenomic information for the respective drug is available. Clicking on a blue icon opens a pop-up containing additional details on known pharmacogenomic effects for this drug. **E) Standard-of-care drugs.** Based on a list of standard-of-care drugs for different subtypes of cancer by the American Cancer Society, drugs approved for the considered cancer subtype are highlighted in green. **F) Target-specific analyses.** Drug-target specific analyses can be started by clicking on the magnifying glass. **G) Side panel.** The collapsible side panel boxes contain help and settings for the analyses. Clicking on a box unfolds its content. Please refer to **Figure A.5** for an overview of the panel content. **H) Help indicator.** Additional information on the content of the side panel boxes is provided when hovering over the respective 'question mark' symbol.

example, the *Filter* box contains various options to filter the results table for drugs with specific pharmacological properties, e.g. to only display antineoplastic agents that act as inhibitors. For an overview of the complete side panel content, please see **Figure A.5**. As additional help, we provide descriptions of the respective modules when hovering over the 'question mark' symbol in each box (**Figure 5.2 H**).

5.2.4.2 Enrichment results

In order to get deeper insights into the regulatory context of a drug target under investigation, we can perform Gene Set Enrichment Analysis within DTI, as described in **Section 5.2.3.4**. For a given drug target, the analysis can be triggered when clicking on the respective *Show KEGG enrichment* button (cf. **Figure 5.2 F**). The enrichment results are then displayed in an additional module in the side panel, see **Figure 5.3 B-D**. In case of a significant enrichment or depletion, the pathway is marked with a red or green arrow, respectively. A visualization of the respective pathways on the KEGG website can be opened by clicking on the $K(EGG)$ button. Details on the enrichment can be investigated by clicking on the *View enrichment results* button (**Figure 5.3 D** and **E**).

5.2.4.3 Visualization of deregulated subnetworks

In order to investigate the regulatory impact that a deregulated drug target might exert on its downstream molecules, DTI provides functionality to compute the most deregulated (or upregulated) subgraph rooted in a drug target of interest. Clicking on the magnifying glass next to the respective drug target and then on the *Compute subnetwork* button (**Figure 5.4 C**) will initiate the analysis. For a better assessment of the subgraph's robustness, not a single subgraph, but several subgraphs of predefined sizes are computed and combined into a consensus graph. Several parameters, including the subgraph size range, can be set in the *Subgraph analysis* side panel (**Figure 5.4 B**), the default range covers subgraphs of sizes three to 15. For the visualization of the resulting subgraphs, we use BiNA, an open-source tool for visualization and analysis of biological networks [514]. BiNA combines sophisticated drawings of biological networks with intuitive mechanisms for navigating, zooming, and searching. DTI uses BiNA Webstart to display the computed subgraphs as a consensus graph, which means that all nodes and all edges contained in the respective subgraphs of different sizes are jointly displayed. The consensus graph is arranged in a hierarchical layout, rooted in the drug target under investigation, for an intuitive visualization of putative signal propagation paths (**Figure 5.4 E**). Nodes are colored according to the deregulation score of the corresponding genes. Moreover, the consensus graph can be interactively investigated and analyzed: users can step through the subgraphs of different sizes to assess whether the graph is extended consistently, which indicates a robust result, or whether there is much variation in the subgraphs. Furthermore, the subgraphs can be annotated with additional information, for example, to highlight mutations or known drug targets (cf. **Figure 5.6**).

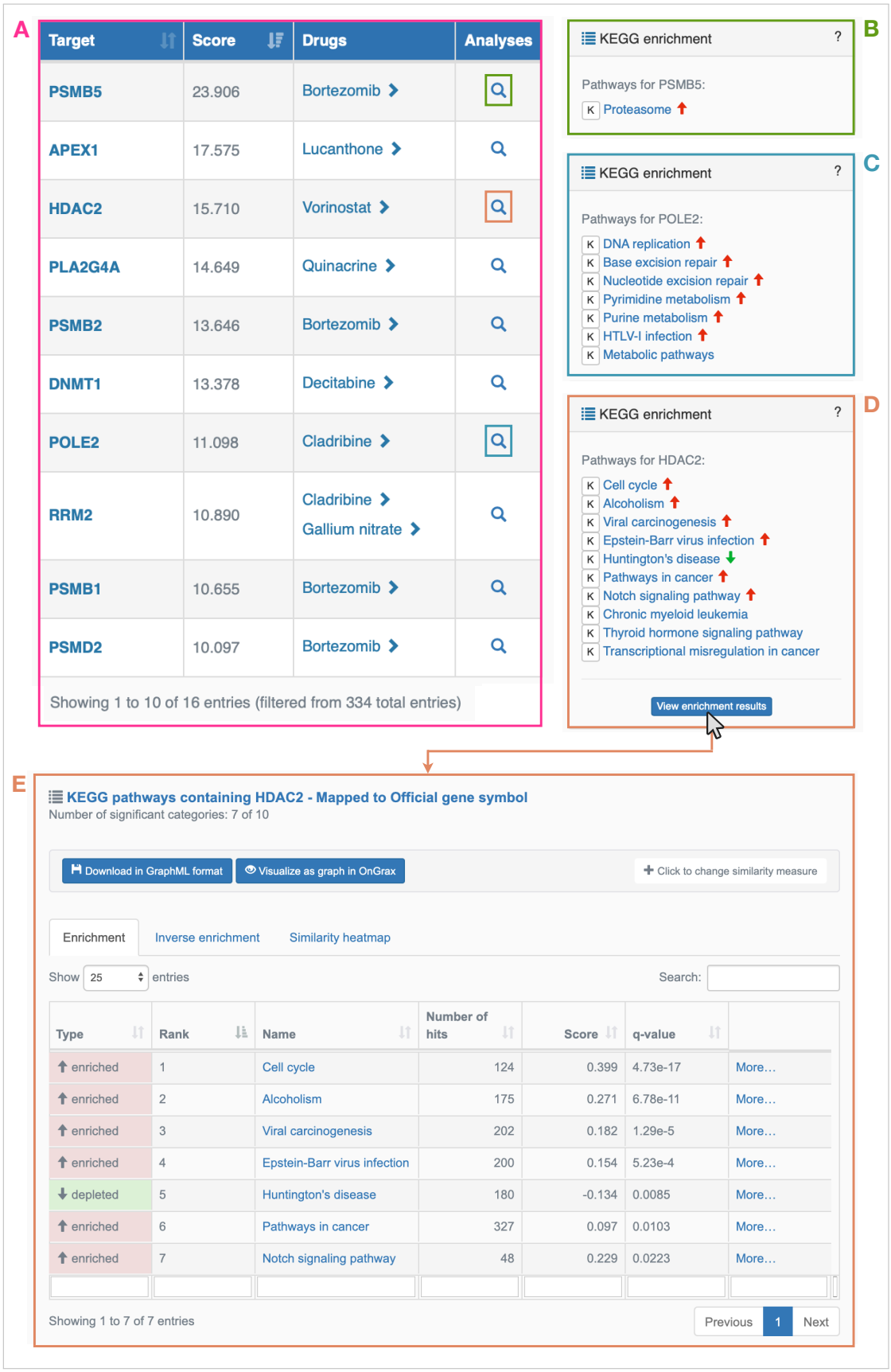


Figure 5.3 Enrichment results in DrugTargetInspector. **A)** Results page for exemplary blastema sample `WS1073TA3` (see **Section 5.3.1**). **B-D)** KEGG enrichment results for the drug targets proteasome 20S subunit beta 5 (PSMB5), DNA polymerase epsilon 2 (POLE2), and histone deacetylase 2 (HDAC2). Red upwards pointing arrows indicate a significant enrichment of the respective pathway in the tumor, while green downwards pointing arrows illustrate a significant depletion. **E)** Detailed view of the HDAC2 enrichment results in GeneTrail2, see **Section 4.2.2** for an in-depth description of the functionality in this results view.

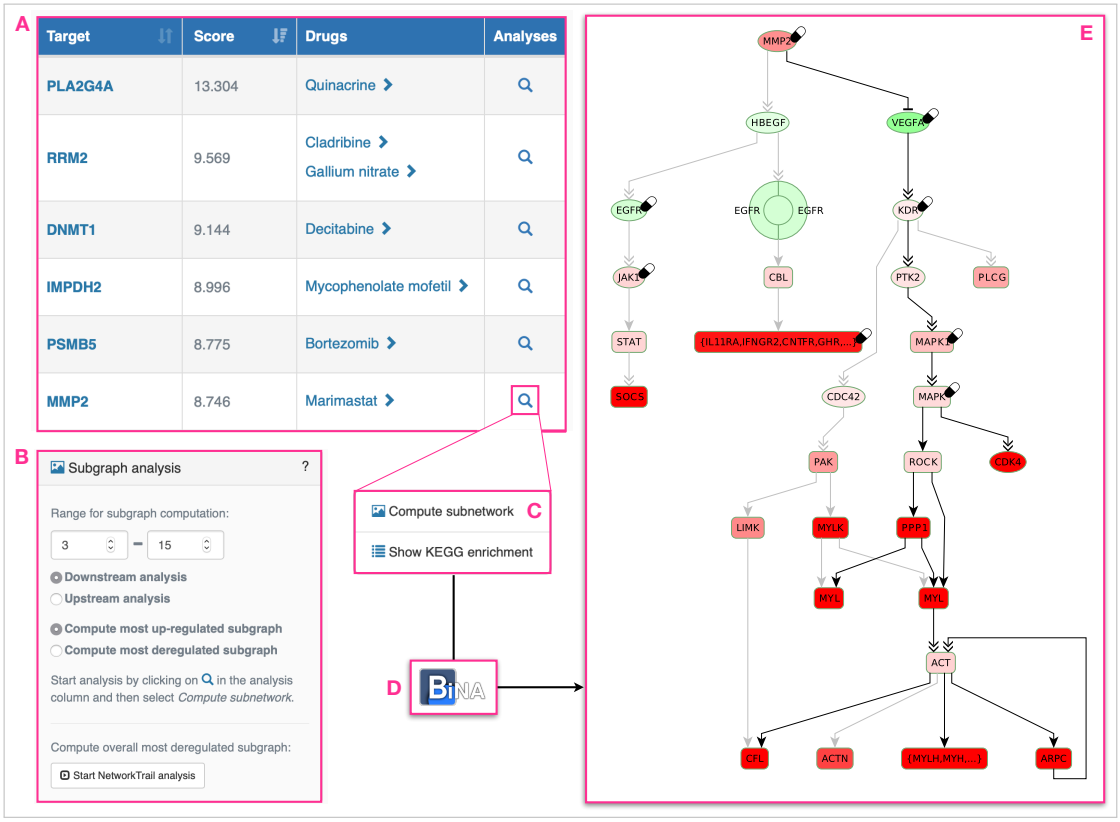


Figure 5.4 Subgraph analysis parameters and results for Wilms tumor sample WS38T in DrugTargetInspector. **A)** Excerpt from DTI results table for Wilms tumor sample WS38T. **B)** Parameter settings for subgraph analysis in the side panel. **C)** *Compute subnetwork* button to initiate the computation of a subgraph rooted in MMP2. **D)** BiNA Webstart is used for visualization of the resulting subnetwork. **E)** Consensus graph of most upregulated subgraphs of sizes between three and 15. The intensity of the node coloring corresponds to the gene's degree of deregulation. Red indicates upregulation and green downregulation. The 'pill' icons flag known drug targets. The node shape describes different types of molecular entities: oval nodes correspond to individual genes/proteins, rounded squares indicate gene families, and circular nodes stand for complexes.

5.3 Case studies

To illustrate the use of DrugTargetInspector, we present clinical case studies addressing three different tumor entities and based on different types of omics data. In **Section 5.3.1**, we will investigate gene expression profiles of Wilms tumors, **Section 5.3.2** presents the results for an integrated analysis of transcriptomics and genomics data in colon adenocarcinoma, and finally, the analysis results for a proteomics data set of a lung adenocarcinoma will be presented in **Section 5.3.3**.

5.3.1 Wilms tumors

Wilms tumors, also known as nephroblastomas, are childhood renal tumors (see also **Section 4.3.4.1**). Although Wilms tumors have a survival rate of more than 90%, there are subtypes associated with a high risk of relapse within the first two years after diagnosis. Wilms tumors are classified into three risk groups: high, intermediate, and low risk according to their risk of relapse. These risk groups are defined by different histologic phenotypes: The blastemal subtype and diffuse anaplasia are considered as high-risk subtypes. The stromal, epithelial, triphasic, regressive, and focal anaplasia subtypes are associated with an intermediate risk, and the completely necrotic subtype shows the lowest risk [552]. Since the tumor biopsies considered in this analysis stem from a European study, chemotherapy was applied before surgery and tumor biopsy. As a standard regimen defined by the Société Internationale d'Oncologie Pediatrique (SIOP) [553], these tumors were treated with a combination of the cytotoxic agents actinomycin D, vincristine, and - in the case of metastases - also doxorubicin. Actinomycin D binds to DNA and inhibits RNA synthesis. Vincristine binds to the microtubular proteins of the mitotic spindle, leading to mitotic arrest or cell death. Doxorubicin is an anthracycline that attacks DNA by several mechanisms: DNA intercalation, strand breakage, and topoisomerase II inhibition. Adjuvant treatment after surgery might be required based on local tumor stage and histological subtype. In the following, we consider Wilms tumor samples of several subtypes and analyze their transcriptomic profiles to identify deregulated drug targets and altered biological pathways that might inform the selection of adjuvant treatment options. For our analysis, we used a gene expression data set of 37 Wilms tumor samples of different subtypes (cf. **Table A.12**). Gene expression experiments were performed using Agilent microarrays and the raw data was processed and normalized using GeneSpring GX [554]. We compared each tumor sample against a set of four normal kidney samples and computed z-scores to assess differential gene expression.

Figure 5.5 shows a heat map of significantly upregulated drug targets per tumor sample. Only genes targetable by known antineoplastic agents are displayed (cf. **Figure A.6** for a complete list of deregulated drug targets). Several observations can be made in this heat map: First, there is substantial heterogeneity, even in samples of the same histological subtype. Second, there are numerous drug targets that are strongly upregulated in many Wilms tumor samples, for example several members of matrix metallopeptidase family, most strikingly MMP16, oftentimes in combination with MMP2 and MMP14. These matrix metalloproteases can all be targeted by the broad spectrum matrix metalloprotease inhibitor marimastat, which might be an interesting treatment option in these cases. Marimastat is an antimetastatic agent and also acts as an angiogenesis inhibitor [427]. Comparing the different pathological subtypes, one can observe

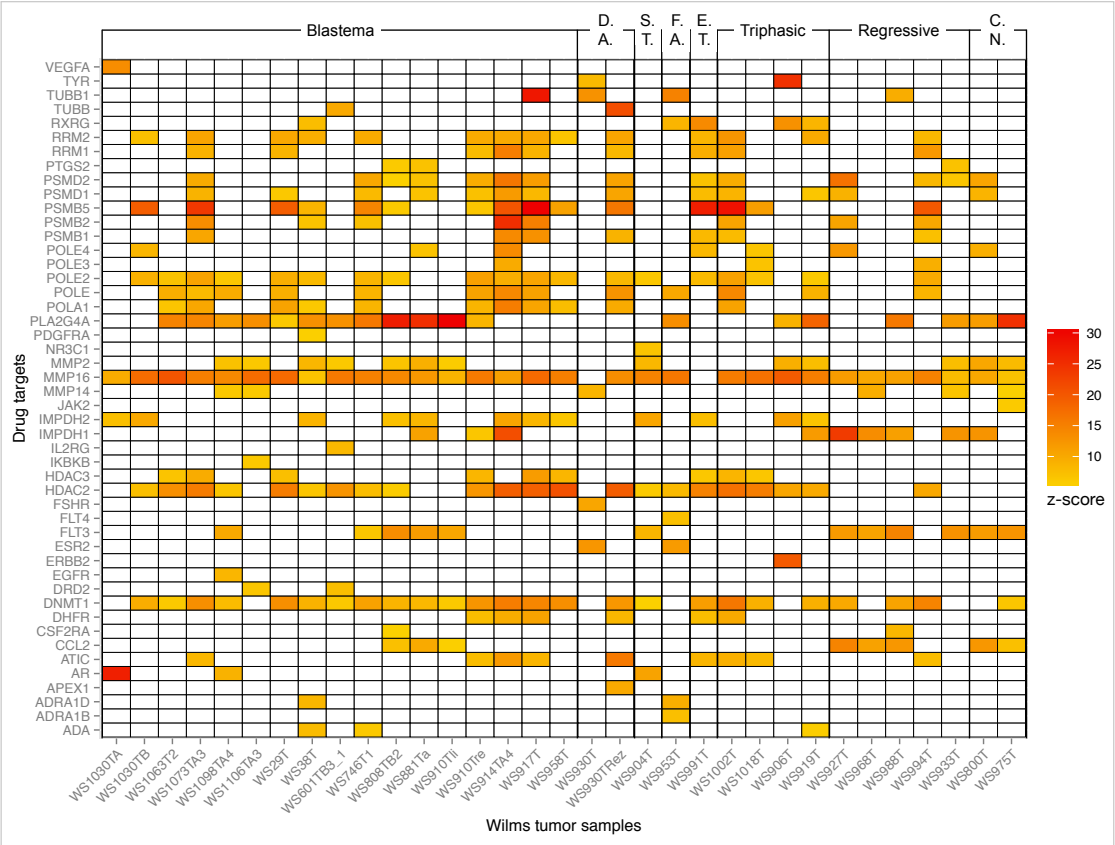


Figure 5.5 Heat map of Wilms tumor samples and a consensus set of deregulated drug targets in DrugTargetInspector. Colored cells in the heat map correspond to significantly upregulated drug targets in the respective samples. White cells indicate that the corresponding drug target was not significantly upregulated in the respective sample. Samples are grouped according to their histological type. D.A.: Diffuse Anaplasia, S.T.: Stromal Type, F.A.: Focal Anaplasia, E.T.: Epithelial Type, C.N.: Completely Necrotic. Only deregulated drug targets that can be targeted by at least one known antineoplastic agent are listed.

that blastema samples generally show higher levels of upregulation for a large number of drug targets than samples of other subtypes. Most of the blastema samples show an over-expression of DNA-methyltransferase 1 (DNMT1) and histone deacetylase 2 (HDAC2). In several samples, HDAC2 upregulation is accompanied by a strong overexpression of several members of the proteasome subunits family (PSMB1, PSMB2, PSMB5, PSMD1, PSMD2). As exemplarily shown for sample `WS1073TA3` in **Figure 5.3**, in such cases, combination treatment with vorinostat and bortezomib could be beneficial, which is also investigated in clinical studies [555]. The phospholipase A2 group IVA (PLA2G4A), which is involved in inflammatory processes, also shows significant upregulation in many blastema samples. Five of the samples (`WS38T`, `WS953T`, `WS991T`, `WS906T`, `WS919T`) exhibit increased levels of RXRG, the retinoid X receptor gamma. This is interesting in terms of elucidation of alternative treatment options, because RXRG can be targeted by vitamin A derivatives such as all-trans retinoic acid (ATRA, tretinoin). Tretinoin acts as cell proliferation inhibitor and differentiation inducer [556]. The positive effect of vitamin A treatment in Wilms tumors has also been described in the literature: Zirn *et al.* [557] and Wegert *et al.* [558] could observe a re-regulation of relapse-associated gene expression patterns after treatment of cultured Wilms tumor cells with ATRA. Also other types of cancer showed to be responsive, in particular acute promyelocytic leukemia [559]. Especially in tumors that are resistant to conventional therapy, ATRA-induced growth inhibition might hence be used as an alternative or additional therapeutic intervention.

5.3.2 Colon adenocarcinoma

In our second case study, we consider colon cancer, the most common type of gastrointestinal cancer. Colon cancer is known to occur spontaneously, but also as a familial cancer [560]. Here, we analyze a data set of colon adenocarcinomas obtained from The Cancer Genome Atlas (TCGA, cf. **Section 3.2.3**), which contains both gene expression measurements and information on somatic mutations [298]. For the computation of deregulated drug targets, we compare samples of tumor tissue against a set of nine healthy tissue controls and calculate z-scores to estimate the changes in gene expression. The corresponding mutation data are converted from the `MAF` file format to the `VCF` file format using the vcf2maf library [561]. In the following, we focus on the sample `TCGA-AA-3542`. In addition to standard cytotoxic drugs, the European Society for Medical Oncology (ESMO) [542] lists the targeted therapeutic agents bevacizumab, regorafenib, cetuximab, and panitumumab as possible treatment options. Two of these drugs, cetuximab and panitumumab, target the epidermal growth factor receptor (EGFR). However, DTI reveals that EGFR is downregulated and additionally mutated, containing a missense variant that is predicted by SIFT to have a deleterious effect (cf. **Section 3.1.2.3**). Moreover, performing a deregulated-subgraph analysis for EGFR shows that most downstream molecules of EGFR are also downregulated, indicating a loss of functionality in EGFR. Thus, treatment with cetuximab and panitumumab might not only be ineffective, but could even cause unwanted side effects.

Figure 5.6 shows that there is a strong over-expression of members of the RAS family, a family of common oncogenes in colon cancer. Hence, in such a case, it might be more effective to add bevacizumab to the regimen, which is also prioritized in DTI over regorafenib. Bevacizumab's molecular target, the vascular endothelial growth factor A (VEGFA), is slightly upregulated, not mutated, and also an upstream molecule of the RAS family. However, DTI also lists other drug

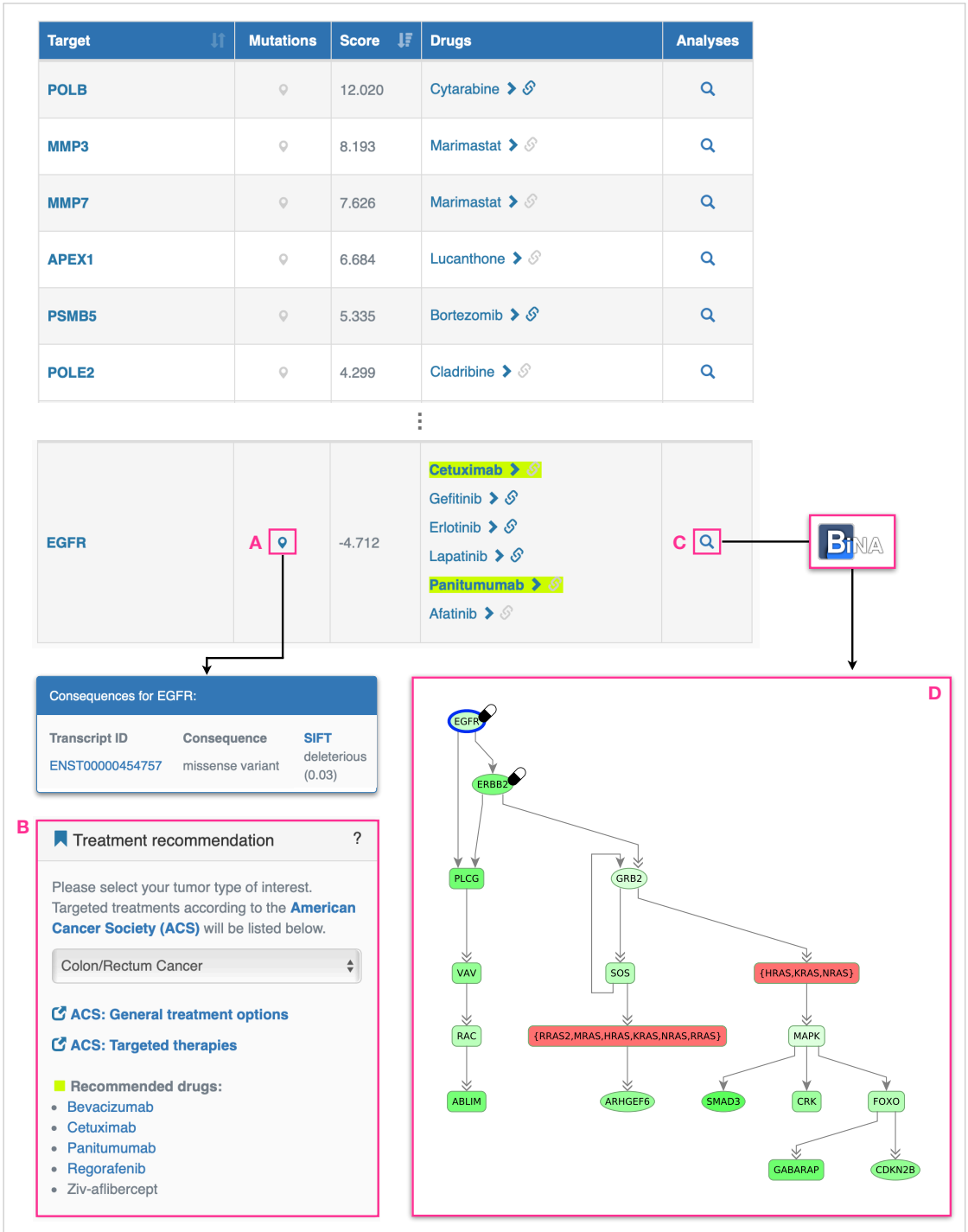


Figure 5.6 **Excerpt from DrugTargetInspector results for colon adenocarcinoma sample TCGA-AA-3542.** **A)** The blue 'map' icon in the *Mutations* column indicates that EGFR contains a mutation. Clicking on this icon opens a pop-up with additional information on the contained mutation(s). **B)** The *Treatment recommendation* side panel contains a list of standard-of-care drugs for the colon/rectum cancer. The corresponding drugs are also highlighted in the results table. **C)** BiNA Webstart is used for visualization of the resulting subnetwork rooted in EGFR. **D)** Consensus graph of most upregulated subgraphs of sizes between 3 and 12. The intensity of the node coloring corresponds to the gene's degree of deregulation. Red indicates upregulation and green downregulation. The 'pill' icons flag known drug targets and the blue border color indicates genes that contain mutations. The node shape describes different types of molecular entities: oval nodes correspond to individual genes/proteins, rounded squares indicate gene families, and circular nodes stand for complexes.

targets that are much stronger upregulated and that might provide alternative treatment options with a putative higher efficacy, as for example members of the matrix metalloprotease family or DNA polymerases, as depicted in **Figure 5.6**. A potential drug regimen targeting several of these highly upregulated proteins could consist of cytarabine, marimastat, and cladribine. This regimen could inhibit the upregulated DNA repair mechanisms (KEGG pathways *Base excision repair* and *Nucleotide excision repair* enriched) that prevent the cells from undergoing apoptosis and promote proliferation. Another approach could be the combination of drugs targeting several distinct deregulated pathways in the tumor.

5.3.3 Lung adenocarcinoma

Lung cancer is the most common cause of cancer death, accounting for about 27% of all cancer deaths [562]. In our third case study, we examine a data set of 18 paired samples of tumor tissue and healthy lung tissue. Half of the tumor samples are adenocarcinoma, the other half squamous cell lung cancer, both of which belong to the class of Non-Small Cell Lung Carcinoma (NSCLC). For each of the samples, we used a Mass Spectrometry (MS)-based workflow to investigate the sample proteomes, see **Section 3.1.3** for a general overview. In contrast to transcriptomics approaches, which use information about the mRNA content of a sample as a proxy for the resulting protein expression, MS-based proteomics allows assessing the protein content directly. Even though current technologies still can only cover relatively abundant proteins, proteomics data offers some distinct advantages compared to transcriptomics data, as the potentially non-linear relationship between mRNA content and protein concentration can severely confound the interpretation of transcriptomics data. In addition, proteomics approaches potentially can detect post-translational modifications that are not detectable in a transcriptomics approach. Here, we used a combination of ion mobility separation with high-performance liquid chromatography and mass spectrometry in a label-free approach, from which we obtained concentrations of more than 3,000 proteins per sample. We computed scores for each sample as the paired difference between protein concentrations in the tumor tissue and the control. Thereby, missing values were set to zero. The resulting score files were uploaded to DrugTargetInspector and the contained UniProt IDs were mapped to the HGNC symbols of their encoding genes. As a specific example, we investigate the paired sample `718`, which contains a specimen of adenocarcinoma tumor tissue (`77-09_718T_AdenoCa1`) and healthy lung tissue from the same patient (`77-10_718_L1`). **Figure 5.7** shows changes in concentration levels for all measured proteins that can be targeted by antineoplastic agents. For NSCLC, ESMO recommends the targeted treatment with EGFR inhibitors (e.g., erlotinib), angiogenesis inhibitors targeting VEGFA (e.g., bevacizumab), and drugs targeting the anaplastic lymphoma receptor tyrosine kinase (ALK), for example crizotinib [563]. As protein abundances for VEGFA and ALK were below the detection threshold and hence not measured, we focus our discussion on EGFR. As can be seen in **Figure 5.7**, the protein EGFR is highly abundant in the tumor sample, while it is even below the detection threshold in the control, which makes it an eligible target for inhibition. Additionally, the FMS-related tyrosine kinase 3 (FLT3), the 5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase (ATIC), and the phosphoribosylglycinamide formyltransferase (GART) are potential drug targets of interest, as for all of them the amount of the corresponding proteins is highly increased in the tumor. For example, FLT3 can be targeted by the tyrosine kinase inhibitors sorafenib and

sunitinib. ATIC and GART are both involved in the de novo purine biosynthetic pathway and can be targeted by pemetrexed. In order to elaborately validate the putative drug targets, of course, information on mutations contained in the tumor and potentially affecting drug targets or important downstream key players would be of interest, as could be seen in the previous case study.

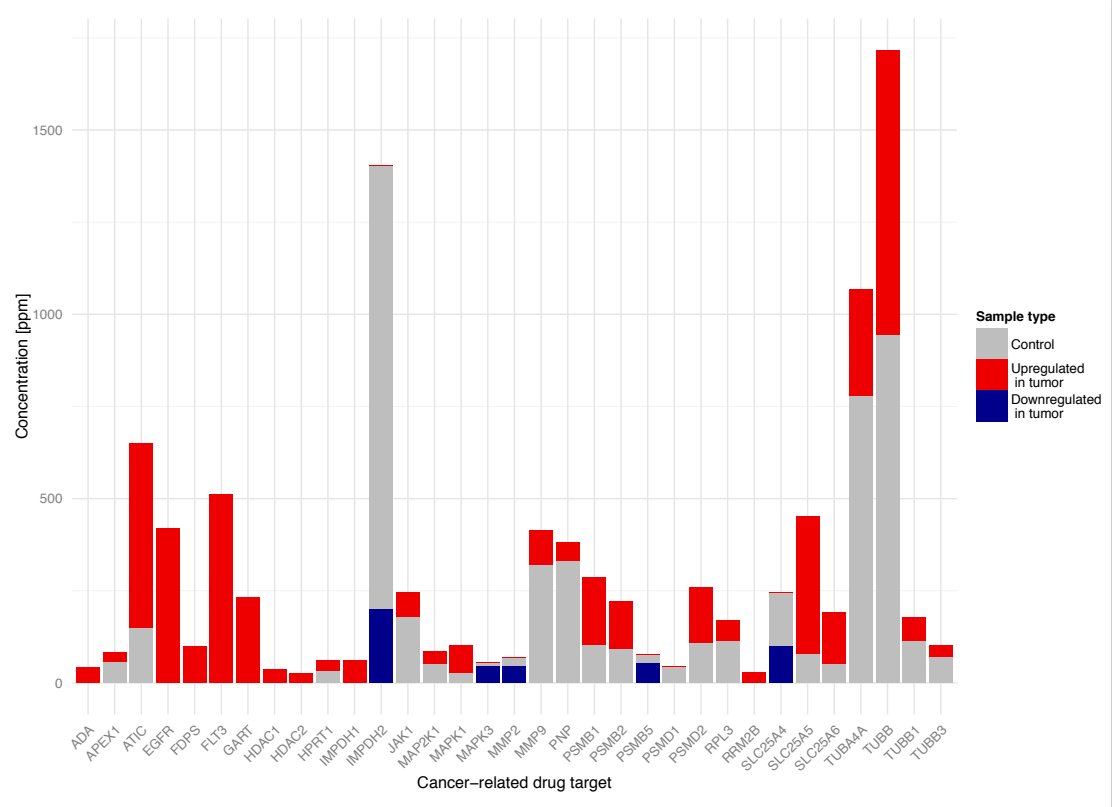


Figure 5.7 Changes in concentration levels for proteins targetable by antineoplastic agents (sample 718). The heights of the single bars indicate absolute concentrations of the corresponding proteins. Gray bars indicate concentration levels in the control. The colored bars show the concentration levels in the tumor, red indicating an increase in abundance, blue a decrease in abundance.

5.4 Discussion

We present DrugTargetInspector, an interactive tool for treatment stratification. DTI analyzes genomics, transcriptomics, and proteomics data sets and provides information on deregulated drug targets, enriched biological pathways, and deregulated subnetworks, as well as mutations and their potential effects on putative drug targets and genes of interest. DTI sorts drug targets based on their degree of upregulation. In order to identify relevant pathways, enrichment analyses based on KEGG pathways can be performed. Using subgraph analyses, users can further investigate which impact the deregulation of a drug target has on the deregulation of other molecules downstream of the drug target in signaling cascades. If this is the case, inhibition and re-regulation of the considered drug target might also positively affect the deregulated downstream processes. DTI can complement the expression-based analyses with genomic variation data: Uploaded `VCF` files are annotated using Ensembl's Variant Effect Predictor in

terms of genomic location and the potential effect on the encoded protein. This additional layer of information can reveal drug targets to be inappropriate for treatment due to, for example, missense mutations as could be seen in our case study on colon adenocarcinoma. As DTI also considers which treatment options are recommended for a certain type of tumor, the tool helps physicians to select the most promising drug or combination of drugs from this resource. However, as DTI's results are not restricted to standard-of-care drugs, other putative drugs can be considered as well in cases where (i) there are no recommendations available, as in high-risk Wilms tumor relapses, or (ii) the effectiveness of standard-of-care drugs is reduced by mutations in the corresponding drug targets. In summary, DTI provides a platform for the integrated analysis of various omics data sets that allows investigating the genomic and transcriptomic properties of a tumor under consideration, which can support physicians in their clinical treatment decision-making process, also elucidating treatment options that might be neglected otherwise.

DrugTargetInspector has been well received by the scientific community and was awarded the *Scientific Excellence Award* at the *Future X Healthcare* conference in 2017, as well as *Landmark in the Land of Ideas* in 2018.

6

ClinOmicsTrail^{bc}

Main parts of this chapter are published in *L. Schneider, T. Kehl, K. Thedinga et al. ClinOmicsTrail^{bc}: a visual analytics tool for breast cancer treatment stratification. Bioinformatics (2019). doi: 10.1093/bioinformatics/btz302.* The ClinOmicsTrail^{bc} web service was developed by myself and Tim Kehl. Hans-Peter Lenhof and I have conceptualized the visualizations and the performed analyses, including the pathway activity measure. The data acquisition, processing, and case studies were performed by me. The neoepitope prediction functionality was developed by Benjamin Schubert and colleagues.

The use of companion diagnostics is more and more becoming an integral part of the personalized treatment of cancer [564] (see also **Section 2.3**). One of the major successes in this field is the assessment of the status of the human epidermal growth factor receptor 2 (HER2/neu, ERBB2), which is overexpressed or amplified in 25% of breast cancer patients. In the HER2-positive case, patients are eligible for treatment with the monoclonal antibody trastuzumab [109]. Still, less than 50% of HER2-positive breast cancers respond to trastuzumab [565]. This observation underlines that the consideration of individual biomarkers does not suffice to capture the complexity of cancer. While the use of gene panels as more comprehensive means of tumor characterization is emerging (cf. **Section 3.1.2.1**), the sole consideration of genomic aberrations only yields a one-dimensional picture of aberrant processes in tumor cells and has been shown to be insufficient to predict treatment response [299, 536, 566]. In contrast, a holistic view on the genomic and molecular dependencies in tumors that utilizes several types of biological information, including (epi-)genomics, transcriptomics, and proteomics data has the potential to lead to actionable and predictive models of cancer [98]. A holistic approach should also encompass the identification and quantification of deregulated biological pathways, as they are the mediators of cancer development and progression and can inform the selection of specific pathway-targeting treatment options.

Considering the continually increasing volumes of quantitative (multi-)omics data on tumors and the vast body of research elucidating molecular and pharmacogenomic dependencies in cancer, clearly illustrates the need for clinical decision support systems. Clinical decision support systems are designed to gather and integrate relevant information and they aim at improving patient outcomes by enabling more confident clinical decisions at the point of care.

To this end, we have developed ClinOmicsTrail^{bc}, a comprehensive visual analytics tool for breast cancer decision support that provides a holistic assessment of standard-of-care targeted drugs, candidates for drug repositioning, and immunotherapeutic approaches. Our tool analyzes and visualizes clinical markers and (epi-)genomics and transcriptomics data sets to identify and evaluate the tumor's main driver mutations, the tumor mutational burden, activity

patterns of core cancer-relevant pathways, drug-specific biomarkers, the status of molecular drug targets, and pharmacogenomic influences. For ClinOmicsTrail^{bc}, we deliberately focused on breast cancer for a proof-of-concept, as breast cancer is the second leading cause of cancer death among women [486]. Moreover, breast tumors, even of the same histopathological subtype, exhibit a high genotypic diversity that impedes therapy stratification and that hence must be accounted for in the treatment decision-making process. Still, albeit ClinOmicsTrail^{bc} is optimized for the analysis of breast cancer data sets, the underlying analysis methods and visualization techniques offered by our web service can also be used for the genetic and molecular characterization of other tumor types by mainly exchanging the tumor-specific underlying databases. In summary, ClinOmicsTrail^{bc} is a powerful integrated visual analytics tool for breast cancer research in general and therapy stratification in particular, assisting oncologists in finding the best possible treatment options for their breast cancer patients based on actionable, evidence-based results. ClinOmicsTrail^{bc} can be freely accessed at `https://clinomicstrail.bioinf.uni-sb.de`.

In the following sections, we will first introduce the concept of molecular tumor boards (**Section 6.1**), followed by a brief overview of breast cancer subtypes and statistics (**Section 6.2**). Related approaches for (breast cancer) treatment stratification will be presented in **Section 6.3**. ClinOmicsTrail^{bc}'s workflow and functionality is the topic of **Section 6.4**. How this functionality can support the treatment decision-making process is then demonstrated using three case studies (**Section 6.5**). Finally, we conclude this chapter with a general discussion of our approach in **Section 6.6**.

6.1 Molecular tumor boards

The challenges of personalized treatment selection for oncology patients have led to the implementation of Molecular Tumor Boards (MTBs) in most cancer centers. MTBs are an extension of traditional multidisciplinary tumor boards in which a group of medical experts, including medical oncologists, surgeons, radiologists, pathologists, and other specialists, decide on the best treatment options for their cancer patients. In addition to the clinical data and treatment history of the patient, MTBs also consider genetic and molecular aberrations contained in the tumor [567]. Typically, molecular tumor boards consider consenting patients who are (i) progressive on all conventional treatment options [568–570] or (ii) who have rare cancers, for which only a few treatments exist [571–573]. Tumor biopsies with a high-enough tumor content (typically $> 20\%$) are analyzed using cancer-specific gene panels, such as those offered by Foundation Medicine [574] or CeGaT [575], sometimes even whole-exome sequencing is conducted [576]. Additional measurements such as transcriptome [571], copy number alterations [119], or the methylome [577] are included occasionally. In addition to potential difficulties in obtaining a tumor biopsy and sufficient quantities of tumor DNA (or RNA) for molecular profiling, the interpretation of the clinical significance of genomic and transcriptomic aberrations is a challenging task [578]. Hence, bioinformatics-driven clinical decision support systems are required to analyze the high-dimensional data sets and to present the treatment-relevant findings in a clear and concise manner to form the basis for the MTB to decide on the treatment of a patient.

6.2 Breast cancer

Breast cancer is the most common type of cancer and the second leading cause of cancer death among women [486] (cf. **Figure 6.1**).

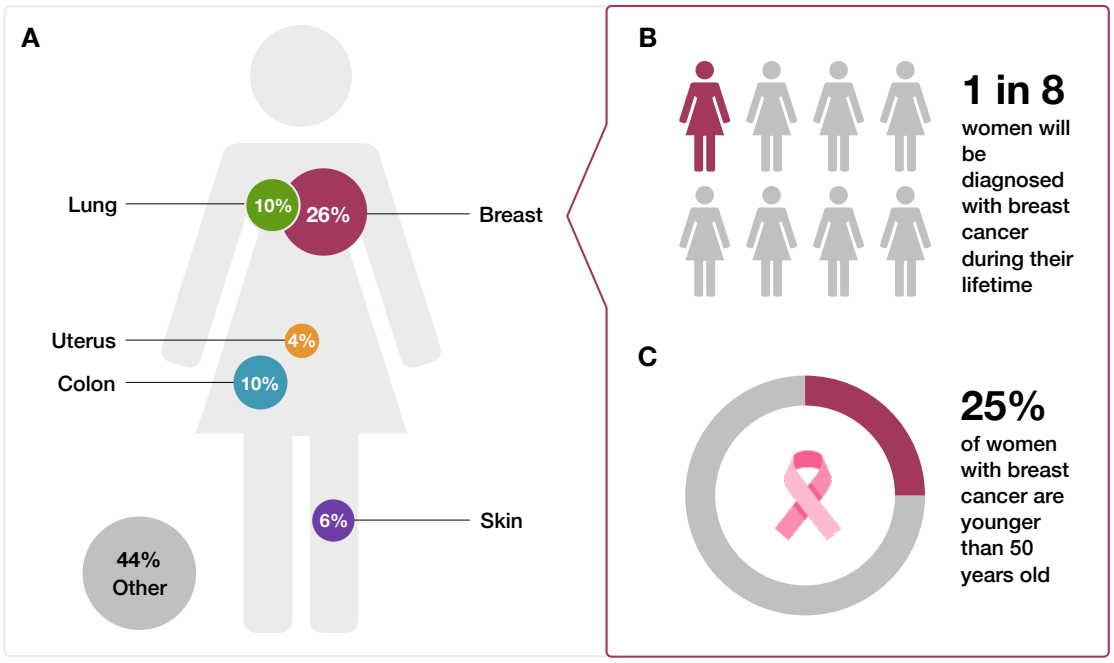


Figure 6.1 **Female cancer statistics**. **A)** Proportion of newly diagnosed female cancer patients by main primary site for a total of 277,940 cases in Germany, 2018 [579]. **B)** Incidence rate for breast cancer in women assuming a lifespan of 80 years [580]. **C)** Data obtained from [581]. The pink ribbon symbol was obtained from [15].

Breast cancer has long been established as a cancer type with several clinically relevant subtypes, which have been identified using both classical immunohistochemistry and gene expression profiling. There are four main molecular subtypes of breast cancer that differ in the composition of relevant receptors and their respective growth rates, but also in their gene expression patterns [490, 582, 583]: (i) breast cancers of type Luminal A are hormone receptor-positive (i.e., they express the estrogen receptor and/or the progesterone receptor), HER2-negative, grow rather slowly, and have the best prognosis. (ii) Luminal B subtypes are hormone receptor-positive, might be HER2-positive, and are slightly more aggressive than the Luminal A subtype. (iii) HER2-enriched tumors are HER2-positive and hormone receptor-negative. The final subtype, (iv) Basal-like or Triple-negative, is the most aggressive one. Its hormone receptor status is negative and HER2 is not amplified. **Figure 6.2** provides an overview of this classification scheme.

The different subtypes show major differences in terms of their incidence, prognosis, and treatment sensitivity [584–586]. Therapeutic approaches to breast cancer include surgery, radiation, and systemic therapy. As the current standard of care, treatment options are typically selected based on a few clinical markers like cancer stage, the presence or absence of hormone receptors, HER2/neu amplification, and the menopausal status of the patient [586, 587]. For the treatment of hormone-dependent (i.e., hormone receptor-positive) breast cancers, endocrine therapy is an integral treatment element [588]. Here, typically the estrogen

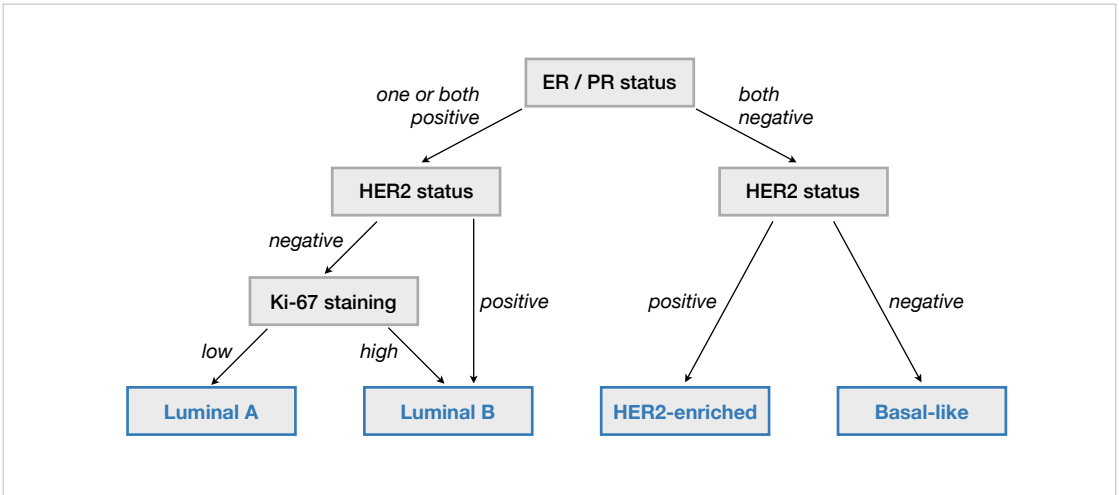


Figure 6.2 Breast cancer subtypes. Classification of breast tumor samples into molecular subtypes (in blue) based on the presence ('positive') or absence ('negative') of the hormone receptors ER and PR (first layer), the overexpression or amplification status of HER2 (second layer) and the tumors growth rates as determined by Ki-67 staining (third layer). ER: estrogen receptor, HER2: human epidermal growth factor receptor 2, PR: progesterone receptor.

receptor inhibitor tamoxifen and, for postmenopausal women, aromatase inhibitors (e.g., exemestane and anastrozole) are used [589]. In the case of HER2-positive breast cancer, there are several targeted drugs available that specifically bind to ERBB2 (HER2/neu) (e.g., trastuzumab, pertuzumab) or to several members of the ERBB family (e.g., lapatinib, neratinib) to prevent dimerization and hence growth signal propagation [590]. In contrast to hormone receptor- and/or HER2-positive breast cancers, the standard of care currently does not provide any targeted treatment options for triple-negative breast cancers, which constitute 10%-20% of all breast cancers [591]. For this aggressive tumor type, the molecular characterization of targetable oncogenic drivers might be especially beneficial [592].

Moreover, studies have shown that differences in the response to treatment, even in patients with the same subtype of breast cancer, correlate with differences in gene expression patterns among these patients, illustrating the urgent need to treat breast cancer with a personalized approach [593].

6.3 Related work

Several tools and web services have been designed to support the breast cancer treatment decision-making process. For example, the web service Adjuvant!Online [594] allows estimating relapse-free and overall survival for women with early breast cancer when treated with adjuvant endocrine therapy or chemotherapy. To this end, clinical parameters like the patient's age at diagnosis and a selection of tumor characteristics like tumor size, grade, and its hormone receptor status are considered. Similarly, the PREDICT web service [595] also provides adjuvant treatment decision support, but additionally considers the HER2 status of the patient. The analysis results in a statistic on the observed relative benefit of four treatment regimens versus no adjuvant treatment in a group of patients with comparable tumor markers. Based on similar clinical input data, CancerMath [596] provides a set of tools to predict life expectancy and to

estimate treatment benefits for various hormonal and chemotherapeutic agents. However, all tools and predictions listed above are solely based on a rather small set of clinical markers, which do not capture the complexity of a tumor [597]. As breast cancer is known for strongly disparate treatment responses, even in patients with common histopathological features [584, 598], it is of utmost importance to additionally consider the tumor's genetic and phenotypic makeup. With the continuous development of biotechnological high-throughput methods, high quality genetic and molecular profiling of tumors has become available at relatively low cost, supporting clinicians in diagnosis, prognosis, and treatment selection. For the clinical assessment of a given breast tumor, nowadays various molecular assays can be used, for example Oncotype DX [106] or MammaPrint [107], which are medium-sized (21 to 70 markers) prognostic gene expression assays that assess the risk for relapse and metastasis. Based on these predictions, the benefit of adjuvant chemotherapy can be determined. Another commonly used multigene signature is the PAM50 classifier, which predicts the molecular subtype of breast cancer samples based on the expression of a set of 50 genes [599]. However, these approaches do not assess the suitability of specific drugs. RecurrenceOnline [600] is an online analysis tool that utilizes microarray-based gene expression data to predict response to hormonal and targeted therapy as well as the risk of recurrence for a given tumor sample. Still, these tests do not consider the effect of mutations that might drive the disease or cause resistance to certain drugs.

Accounting for the influence of mutations on treatment response, there are also several companies (e.g., CeGaT [575] or Foundation Medicine [574]) that commercially offer clinical decision support based on next-generation sequencing and proprietary genomic profiling assays. The provided reports contain information on pathogenic mutations with known pharmacogenomic effects on cancer drugs. While this information is already very valuable, the underlying analyses only focus on the investigation of genomic aberrations, neglecting all other layers of cellular regulation and disease manifestation.

In order to obtain a more holistic picture of aberrant processes in tumors, we have developed DrugTargetInspector (see **Chapter 5**), an interactive assistance tool for the investigation of differentially expressed or mutated molecular targets of known drugs. Based on this, approved drugs for different cancer types and putative drug repositioning candidates can be assessed. Especially the aspect of considering aberrant pathways to identify biomarkers and to inform treatment decision-making has been well adopted in the research community: Alcaraz *et al.* have developed PathClass [601], a web service that predicts molecular subtypes of breast cancer samples based on gene expression data and using different methods including PAM50 [599], SCMGENE [602], SCMOD1/2 [603, 604], and KeyPathwayMiner [605] (cf. **Section 3.3.3.3**). Similarly, Lee *et al.* use the metagene-based method of COndition-Responsive Genes (CORGs, cf. **Section 3.3.3.2**) on differentially expressed genes, copy number variations, and miRNA target genes to predict molecular subtypes of breast cancer using k-means clustering [606]. For the identification of drug repositioning candidates for the treatment of breast cancer, Mejia-Pedroza *et al.* derive pathway activities from Pathifier (cf. **Section 3.3.3.2**) to identify pathway-based subtypes potentially susceptible to treatment with drugs targeting these aberrant pathways [607].

In order to translate multi-omics, holistic analyses of tumor samples into clinical practice, clinical markers have to be analyzed jointly with large and complex omics data sets. This, in turn, requires easy-to-use bioinformatics tools able to integrate different molecular and

clinical data sets and to extract the most relevant information. On top of this, the determined tumor characteristics have to be evaluated with respect to a comprehensive body of medical, pharmacological, and biological knowledge to gain actionable insights from the data. The obtained results finally have to be presented in a concise manner that facilitates interpretation. Here, we present ClinOmicsTrail^{bc}, an interactive visual analytics tool for breast cancer treatment stratification. ClinOmicsTrail^{bc} offers rich functionality for the integration and analysis of clinical markers as well as transcriptomics and (epi-)genomics data sets with respect to a broad spectrum of biological, pharmacological, and medical knowledge. Our tool provides a comprehensive assessment of a variety of treatment options based on the tumor's main driver mutations, the overall tumor mutational burden, activity patterns of core breast cancer-relevant pathways, drug-specific predictive biomarkers, the status of molecular drug targets and pharmacogenomic implications. **Table 6.1** shows how ClinOmicsTrail^{bc} excels with respect to breadth and depth of analyses compared to related approaches.

	Adjuvant therapy assessment	Survival time prediction	On-label drug assessment	Off-label drug assessment	Immunotherapy assessment	Pathological markers as input	Genomics data as input	Epigenomics data as input	Transcriptomics data as input	Interactive results provided
Adjuvant!Online [608]	✓	✓	(✓)	✗	✗	✓	✗	✗	✗	✗
PREDICT [595]	✓	✓	(✓)	✗	✗	✓	✗	✗	✗	✗
CancerMath [596]	(✓)	✓	(✓)	✗	✗	✓	✗	✗	✗	✗
Oncotype DX [106]	✓	✗	(✓)	✗	✗	✓	✗	✗	✓	✗
MammaPrint [107]	✓	✗	✗	✗	✗	✓	✗	✗	✓	✗
DrugTargetInspector [609]	✗	✗	✓	✓	✗	✗	✓	✗	✓	✓
CeGaT [575]	✗	✗	(✓)	✗	✓	✗	✓	✗	✗	✗
FoundationOne CDx [574]	✗	✗	✓	✓	✓	✗	✓	✗	✗	✗
ClinOmicsTrail^{bc} [610]	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓

Table 6.1 **Comparison of different tools for clinical breast cancer decision support**. The tools are assessed in terms of their considered input data types, the performed analyses / predictions, and the presentation of the results. The green checkmark indicates that a certain feature is given, the blue checkmark in parentheses stands for a limited extent to which a feature is provided / considered, and the red cross means that the considered feature is not provided.

6.4 Workflow and functionality

ClinOmicsTrail^{bc} guides breast cancer therapy selection by evaluating available therapeutic regimens in the context of the individual molecular tumor characteristics. To this end,

ClinOmicsTrail^{bc} analyzes and integrates clinical, genomics, and transcriptomics data. These tumor characteristics are then combined with *a priori* knowledge from clinical practice guidelines as well as from various medical, pharmacological, and biological databases to assess therapy options, both on-label and off-label. ClinOmicsTrail^{bc} reports these results in an intuitive and interactive manner, highlighting characteristics that might support or contraindicate the use of specific therapy options. These characteristics include a variety of relevant factors for treatment success (e.g., biomarkers, molecular drug targets, and drug metabolism) as well as breast cancer-driving pathways. **Figure 6.3** gives an overview of the ClinOmicsTrail^{bc} workflow, covering all steps from sample data input, over preprocessing and analysis steps, to the visualization of the results. The respective workflow components will be described in more detail in the following sections.

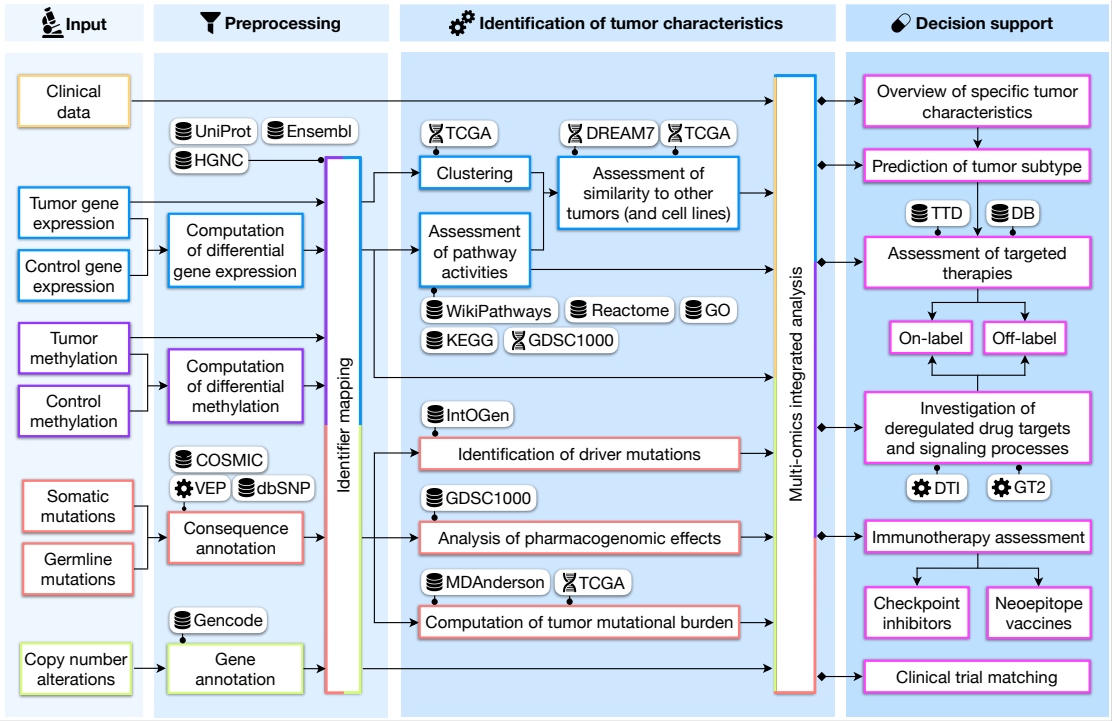


Figure 6.3 **Overview of ClinOmicsTrail^{bc} workflow**. The box border colors correspond to the type of data used in the corresponding step or analysis. Orange: clinical data, blue: gene expression data, purple: methylation data, red: mutation data, green: copy number alterations. The pink border color in the last column stands for results and potential analysis endpoints. Integrated databases are indicated by a database icon, third party tools by a gear wheel, and molecular data sets by the double-helix symbol. COSMIC: Catalogue Of Somatic Mutations In Cancer, DB: DrugBank, DREAM7: Dialogue for Reverse Engineering Assessments and Methods – drug sensitivity prediction challenge, DTI: DrugTargetInspector, GDSC1000: Genomics of Drug Sensitivity in Cancer, GO: Gene Ontology, GT2: GeneTrail2, HGNC: HUGO Gene Nomenclature Committee, IntOGen: Integrative Onco Genomics, KEGG: Kyoto Encyclopedia of Genes and Genomes, MDAnderson: MD Anderson Cancer Center, TCGA: The Cancer Genome Atlas, TTD: Therapeutic Target Database, VEP: Variant Effect Predictor. The displayed icons were obtained from [15].

6.4.1 Tumor-specific input data and preprocessing

ClinOmicsTrail^{bc} allows the user to integrate clinical data of the tumor under investigation with its (epi-)genetic alterations and gene expression measurements (cf. **Figure 6.3**, first and second

column). Details on the considered types of input data and their relevance for the treatment decision-making process will be provided in the following sections.

6.4.1.1 Clinical data

ClinOmicsTrail^{bc} analyzes the status of several standard clinical markers for breast cancer diagnosis and treatment. Firstly, the status of the estrogen receptor (ER) and the progesterone receptor (PR) are considered as they play a crucial role in breast cancer development and progression. Their presence or lack – in combination with the menopausal status of the patient – informs the eligibility of several types of drugs, including aromatase inhibitors and estrogen receptor-targeting drugs. Another critical biomarker to be considered is the status of HER2 [611]. An overrepresentation of this receptor on tumor cells makes them sensitive to treatment with, for example, antibody-based inhibitors like trastuzumab or pertuzumab. Also, information on tumor growth (Ki-67 staining, s-phase fraction), the histopathological subtype, tumor size and grade, lymph node and metastasis status, as well as clinical metadata like a patient ID, the origin of the sample (primary tumor vs. metastasis), the fraction of tumor tissue in the sample, and the date of biopsy can be provided to ClinOmicsTrail^{bc}.

6.4.1.2 Molecular and genetic data

Besides the clinical biomarkers, various types of molecular data, covering several layers of genomic and transcriptomic regulation, can be processed and investigated.

Based on gene expression measurements of a tumor sample and a matched control, ClinOmicsTrail^{bc} computes for each gene a score mirroring its differential gene expression. These scores are used - amongst others - to estimate the activity of breast cancer-relevant signaling pathways (cf. **Section 6.4.2**). Gene expression data can be provided as a whitespace-delimited score file or gene expression matrix. For the latter, a single sample of interest has to be selected as well as one or several non-tumor control samples. Based on the selection, either a log fold change or a z-score is computed to assess differential gene expression.

For the analysis of genetic variations, ClinOmicsTrail^{bc} requires (whole genome or exome) mutation data of the tumor sample and a (e.g., blood-derived) control to identify somatic and germline mutations. These will be used for the identification and prioritization of driver mutations, the assessment of pharmacogenomic effects, and the computation of tumor mutational burden (cf. **Section 6.4.2**). Mutation data can be uploaded to ClinOmicsTrail^{bc} in Variant Call Format (`VCF`, cf. **Section A.1.4**). Ideally, the provided `VCF` file should contain information on the tumor sample and a matched control as to clearly differentiate between somatic and germline mutations. In cases where only a tumor sample is provided, ClinOmicsTrail^{bc} estimates, based on allele frequency, whether a mutation is likely to be germline or somatic. In order to assess the impact of protein-coding mutations on the disease phenotype, all mutations are annotated with their effects on the corresponding proteins (e.g., missense variant, stop gain, frameshift) using Ensembl's Variant Effect Predictor (VEP) [220]. Additionally, the contained mutations are cross-referenced with dbSNP [226] and COSMIC [612] for further details on the potential functional impact of the mutation.

As altered copy numbers are believed to account for up to 85% of dysregulated gene expression in breast cancers [613], ClinOmicsTrail^{bc} also considers this type of data for the identification

of driver genes and the holistic assessment of altered processes in the tumor. Copy number alterations can be uploaded in segmented data file format (`SEG`, cf. **Section A.1.6**) as log-ratios of tumor copy numbers in relation to normal copy number levels. The copy number alterations of the contained genomic regions are mapped to genes using the reference genomes GRCh37/38 [398] and gene annotations from Gencode [399].

Besides the aforementioned genomic aberrations, also epigenomic changes (e.g., differential methylation of promotors) can contribute to tumor initiation and progression. Breast cancer is typically characterized by a combination of global hypomethylation and local hypermethylation, where the latter is likely to silence growth-regulatory genes [614]. Methylation data can be provided to ClinOmicsTrail^{bc} as white-space separated files that contain methylation scores (e.g., beta values of promoter methylation) per gene identifier.

For all omics data types, we support all commonly used identifier types, including HGNC gene symbols [272], NCBI EntrezGene IDs [270], and UniProt identifiers [274]. The identifiers used in the provided omics data sets are unified in an automatic identifier mapping step for seamless integration in the following analyses.

Since in a clinical setting not all of the described clinical, genomics, and transcriptomics data might be available, ClinOmicsTrail^{bc} only requires the user to provide gene expression data. However, the more types of omics data are available, the better ClinOmicsTrail^{bc} can reveal its full potential, providing additional analyses and a more comprehensive assessment of tumor characteristics. A summary of all uploaded data sets and the provided clinical information is given on the results page for further reference.

6.4.2 Identification of tumor characteristics

Based on the various types of clinical and molecular data described in **Section 6.4.1**, ClinOmicsTrail^{bc} performs a variety of integrated analyses (cf. **Figure 6.3**, third column) with a focus on breast cancer-relevant driver genes and signaling pathways.

In a first step, ClinOmicsTrail^{bc} identifies tumor-driving (epi-)genomic aberrations, including mutations, copy number variations, and DNA methylation (**Section 6.4.2.1**). As these alterations also manifest in the activities of signaling cascades, we compute pathway activities for a set of 20 core cancer-associated pathways based on the differential gene expression of the involved genes (**Section 6.4.2.2**). Finally, based on the complete expression profile, ClinOmicsTrail^{bc} provides a clustering with respect to more than 500 breast cancer profiles from TCGA that allows to assess the tumor's intrinsic subtype (**Section 6.4.2.3**).

6.4.2.1 Identification of tumor-driving (epi-)genomic aberrations

Tumors can potentially contain a plethora of mutations that are usually divided into driver and passenger mutations based on their impact on disease development. Driver mutations are thereby defined as those that confer a selective growth advantage to the cell [615]. Genes commonly affected by driver mutations in breast cancer are, e.g., TP53 (mutated in 33% of cases) [616], PIK3CA (30%) [617], or GATA3 (9%) [618]. For the identification and prioritization of tumor-specific driver genes, ClinOmicsTrail^{bc} uses the IntOGen database [619], which identifies driver genes based on their predicted role in tumorigenesis and mutation frequency in large tumor cohorts. ClinOmicsTrail^{bc} provides a prioritized list (sorted by frequency in breast cancer)

of putative driver genes contained in the tumor. For each of these driver genes, details about the contained mutations and their potential impact on the protein, as well as copy number and gene expression scores are shown. The *Mutation status* column contains color-coded symbols indicating the predicted severity of a mutation. Clicking on the corresponding symbol opens additional details on the contained mutation(s), including links to COSMIC and dbSNP for known variants. Additionally, the prioritized table can be extended by the remaining mutations for a complete assessment of the tumor's genetic aberrations.

Mutations contained in a tumor may not only alter the functioning of the corresponding proteins and affect pathway activities, but they can also modulate a tumor's response to drugs with respect to efficacy and toxicity [620]. A prominent example of a pharmacogenomic effect is the treatment of colon cancer with cetuximab that becomes ineffective in the presence of an activating mutation in the kirsten rat sarcoma viral oncogene homolog (KRAS) [8]. Besides somatic mutations, also germline alterations can affect drug sensitivity. Many anticancer drugs need to be metabolized into their active forms by enzymes in the liver. The estrogen receptor-targeting drug tamoxifen, for example, requires cytochrome P450 2D6 (CYP2D6) for activation. However, in cases where the enzymatic activity of CYP2D6 is restricted by a mutation or its expression is reduced by a chromosomal deletion, a treatment with tamoxifen may be ineffective [621]. In order to account for such pharmacogenomic effects, ClinOmicsTrail^{bc} investigates the genomic and transcriptomic state of relevant drug-processing enzymes and resistance-promoting factors to evaluate the applicability of drugs or the potentially required adaption in dosage for the considered case (cf. **Section 6.4.3**). Additionally, for a given mutation, ClinOmicsTrail^{bc} displays the putative impact of this mutation on a variety of drugs as predicted by the Genomics of Drug Sensitivity in Cancer (GDSC1000) database [533].

6.4.2.2 Assessment of pathway activities

Tumors are driven by the aberrant activity of key signaling pathways that, e.g., promote tumor growth or hinder apoptosis [622]. Identifying the involved pathways and quantifying their deregulation is a major step towards the understanding of the underlying malignancy processes. In order to obtain an overview of altered processes in a breast tumor, we consider pathway activities of a set of 20 core breast cancer-relevant pathways [623–625]. These activity patterns can, in turn, be used to assess characteristics of tumor subtypes and to inform a treatment decision. For example, tumors with specifically high activities in PIK-AKT-mTOR signaling could potentially profit from treatment with an AKT inhibitor (e.g., ipatasertib [626]) or an mTOR inhibitor (e.g., everolimus [627]).

Computation of pathway activities: ClinOmicsTrail^{bc} offers functionality to assess pathway activities for a set of 20 core breast cancer-relevant pathways. We compute the activity of a pathway i based on the deregulation scores of the pathway genes Γ_i, weighted by their relevance $\mathbf{w_i}$ for the respective pathway's activity. **Algorithm 6.1** provides an overview of the computation in pseudocode. As different databases provide different gene sets for the respective pathways, we merge the corresponding gene sets from KEGG [282], GO [424], Reactome [283], and WikiPathways [628] to obtain a more comprehensive set of genes as the basis for the pathway activity computation (cf. **Section A.8.1.1**).

We hypothesize that targeted drugs are especially effective in cases where their target pathway is highly active and alternative cancer-driving pathways are not. We take advantage of the assumed relationship between a pathway's activity and a corresponding drug's efficacy to compute the weights $\mathbf{w_i}$. To this end, we consider all 49 breast cancer cell lines from the Genomics of Drug Sensitivity in Cancer (GDSC1000) database and their sensitivities for a large panel of drugs targeting various pathways [533]. The authors provided drug sensitivity scores as IC50 values, i.e. the log-transformed concentration of an inhibitor that decreases the biotransformation rate of its target's substrates by 50%. For a given pathway of interest i, we select the set D_i of drugs from GDSC1000 that target this pathway. For each of those drugs $d_{ij} \in D_i$, we compute Pearson's correlation between the drug's IC50 values and the gene expression measurements across cell lines. For each of those correlation coefficients c_{jk}, we also compute a p-value assessing the significance of its deviation from zero (p_{jk}). As we assume that pathway-activating genes have a negative correlation with IC50, we switch the sign of c_{jk} by multiplying with -1 to make the final scores more intuitive. This results in two matrices of dimensions $l_i \times m_i$ each, where l_i corresponds to the size of the gene set Γ_i and m_i to the size of the drug set D_i. The matrices are transformed as follows: the correlation coefficients are z-transformed per drug and then averaged cross drugs yielding a list of correlation-based scores per pathway $\hat{\mathbf{c}}_\mathbf{i}$. The p-values are aggregated across drugs using Fisher's method [300]. The aggregated p-values are then -log10-transformed to obtain scores per pathway and gene $\mathbf{p}_\mathbf{i}^*$. The larger the score p_{ik}^* for a gene k, the more relevant it is as an indicator for the pathway's activity. As the scores in $\mathbf{p}_\mathbf{i}^*$ are all positive, we recover the direction of the gene's effect, i.e. whether it acts as an activator or repressor of the pathway, from the sign of the corresponding correlation-based score $\hat{c}_{ik}$. The final weights $\mathbf{w_i}$ are computed using the Hadamard product [629]: $\mathbf{w_i} = sgn(\hat{\mathbf{c}}_\mathbf{i}) \odot \mathbf{p}_\mathbf{i}^*$.
The pathway activity $\phi_i(\mathbf{t})$ for a tumor sample $\mathbf{t}$ and a pathway i is computed as $\phi_i(\mathbf{t}) = \mathbf{w_i} \cdot \mathbf{t_i}$, where $\mathbf{t_i}$ contains the deregulation scores for the subset of genes in the gene set Γ_i. Pathway activities per pathway are finally embedded into a range between 0 and 1.

In order to assess the significance of computed pathway activities for a sample under investigation, empirical p-values can be derived from permutation testing. Based on a user-defined number of permutations, the scores of differential gene expression are randomly permuted and the corresponding pathway activities are re-computed. One-sided p-values for the sample's (actual) pathway activities are computed relative to the mean of the empirical background distribution, where a right-sided p-value is computed if the sample's pathway activity for a specific pathway is larger than the mean of the background distribution and a left-sided p-value if it is smaller. In a last step, the derived p-values are adjusted for multiple hypothesis testing using the Benjamini-Hochberg method [310].

Input: Normalized gene expression matrix $\mathbf{E} \in \mathbb{R}^{n \times p}$ for n cell lines and p genes, matrix $\mathbf{S} \in \mathbb{R}^{n \times d}$ of IC50 drug sensitivites for n cell lines and d drugs, lists of genes belonging to each of the 20 considered pathways i: $\Gamma_i = \{g_{i1}, ..., g_{il_i}\}$, lists for drugs targeting each of the 20 considered pathways i: $D_i = \{d_{i1}, ..., d_{im_i}\}$, normalized gene expression vector $\mathbf{t} \in \mathbb{R}^p$ for tumor sample under investigation

Output: Pathway activities $\phi_i(\mathbf{t})$ for each of the 20 considered pathways i in tumor sample $\mathbf{t}$

foreach *pathway* i *in* $\{1, ..., 20\}$ **do**

$\mathbf{E_i} \leftarrow \mathbf{E}[; g_{i1}, ..., g_{il_i}]$ //Subset columns of $\mathbf{E}$ for pathway genes Γ_i
$\mathbf{S_i} \leftarrow \mathbf{S}[; d_{i1}, ..., d_{im_i}]$ //Subset columns of $\mathbf{S}$ for pathway-targeting drugs D_i
$\mathbf{C_i} \leftarrow$ matrix(nrow=l_i, ncol=m_i) //Matrices to store correlation values
$\mathbf{P_i} \leftarrow$ matrix(nrow=l_i, ncol=m_i) //and corresponding p-values
for j *in* $\{1, ..., l_i\}$ **do**
  for k *in* $\{1, .., m_i\}$ **do**
    $(c_{jk}, p_{jk}) \leftarrow$ computeCorrelationAndPvalue($\mathbf{E_i}[; g_{ij}]$,$\mathbf{S_i}[; d_{ik}]$)
    $\mathbf{C_i}[j; k] \leftarrow -1 \cdot c_{jk}$
    $\mathbf{P_i}[j; k] \leftarrow p_{jk}$
  end
end

//Column-wise z-transform $\mathbf{C_i}$ into $\hat{\mathbf{C}}_\mathbf{i}$
$\hat{\mathbf{C}}_\mathbf{i} \leftarrow$ matrix(nrow=l_i, ncol=m_i)
for j *in* $\{1, ..., l_i\}$ **do**
  for k *in* $\{1, .., m_i\}$ **do**
    $\hat{\mathbf{C}}_\mathbf{i}[j; k] \leftarrow \frac{\mathbf{C_i}[j;k] - \mu_k}{\sigma_k}$
  end
end
$\hat{\mathbf{c}}_\mathbf{i} \leftarrow$ takeRowAverages($\hat{\mathbf{C}}_\mathbf{i}$) //Take row-wise average
$\mathbf{p_i} \leftarrow$ aggregateRowPvaluesUsingFisher($\mathbf{P_i}$)//Row-wise aggregation
$\mathbf{p_i^*} \leftarrow -1 \cdot \log_{10}(\mathbf{p_i})$
$\mathbf{w_i} \leftarrow sgn(\hat{\mathbf{c}}_\mathbf{i}) \odot \mathbf{p_i^*}$
$\mathbf{t_i} \leftarrow \mathbf{t}[g_{i1}, ..., g_{il_i}]$ //Subset expression vector $\mathbf{t}$ for pathway genes Γ_i
$\phi_i(\mathbf{t}) \leftarrow \mathbf{w_i} \cdot \mathbf{t_i}$
$\phi_i(\mathbf{t}) \leftarrow$ normalizeToRangeFrom0To1($\phi_i(\mathbf{t})$)
end

Algorithm 6.1 Pseudocode for computation of pathway activities. Computation of pathway activities for 20 breast cancer-releveant pathways and a tumor sample t. The input data $\mathbf{E}$, $\mathbf{S}$, and D_i are derived from GDSC1000 [533]. The variables μ_k and σ_k stand for the sample mean and sample variance of the k-th column in C_i, respectively. P-value aggregation is performed using Fisher's method [300]. The symbol '$\odot$' indicates the Hadamard product of compontent-wise vector multiplication [629].

The computed pathway activities are displayed in a radar chart, where each radar axis represents the activity of a pathway (cf. **Figure 6.4**). As a reference, the pathway activity patterns of more than 500 primary tumor samples from the TCGA breast cancer cohort and 45 breast cancer cell lines [299] can be interactively added to the plot by clicking on the corresponding sample's checkbox (cf. right panel in **Figure 6.4**). The reference samples are sorted in decreasing order of similarity to the sample under consideration. The similarity is assessed based on the mean-squared distance of corresponding pathway activity scores. Clicking on the sample name yields information on clinical markers and in the case of cell lines additionally details on drug sensitivities and growth rates (cf. **Figure A.17**).

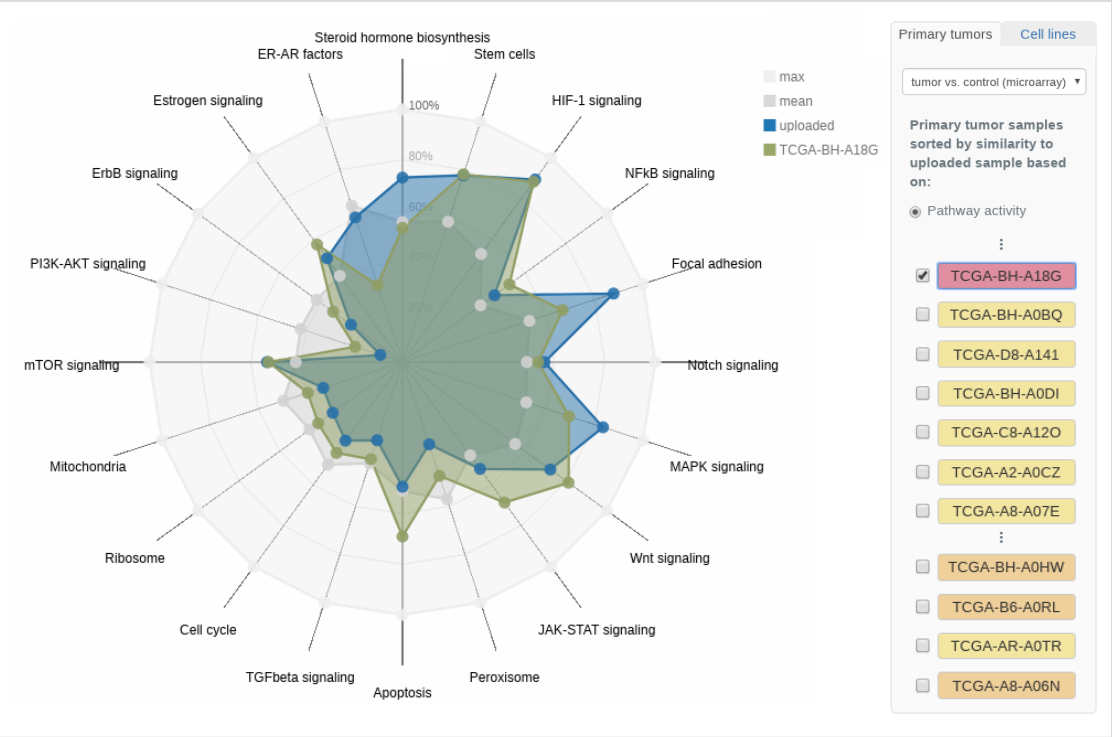


Figure 6.4 **Radar chart of pathway activities in ClinOmicsTrail[bc]**. The pathway activities of a set of 20 core breast cancer pathways for the user-provided tumor sample (`TCGA-AN-A0XN`, **Section 6.5.1**) are colored in blue. Reference samples from TCGA as well as breast cancer cell lines can be added to the visualization interactively. Here, the triple-negative `TCGA-BH-A18G` (in green) shows a similar activity pattern to the sample under investigation. The molecular subtype of the respective reference samples is color-coded in the side panel on the right: basal-like - red, luminal A - yellow, luminal B – orange.

6.4.2.3 Clustering

There are four main molecular subtypes of breast cancer (cf. **Section 6.2**) that differ in the composition of relevant receptors and their respective growth rates, but also in their gene expression patterns [490]. In order to investigate a sample's intrinsic subtype, we compute a clustering of the sample under investigation in comparison to more than 500 breast tumor samples obtained from TCGA [630]. To this end, we use the classic Principal Component Analysis (PCA) [631] as well as t-distributed Stochastic Neighbor Embedding (t-SNE) [632], a non-linear dimension reduction technique that captures the similarity of samples in a two-dimensional space. In the resulting visualization, TCGA's samples are color-coded according

to their molecular subtypes [298] (cf. **Figure 6.5**). The location of the uploaded sample within this visualization describes its molecular similarity to the TCGA samples and can hence provide additional evidence for its subtype.

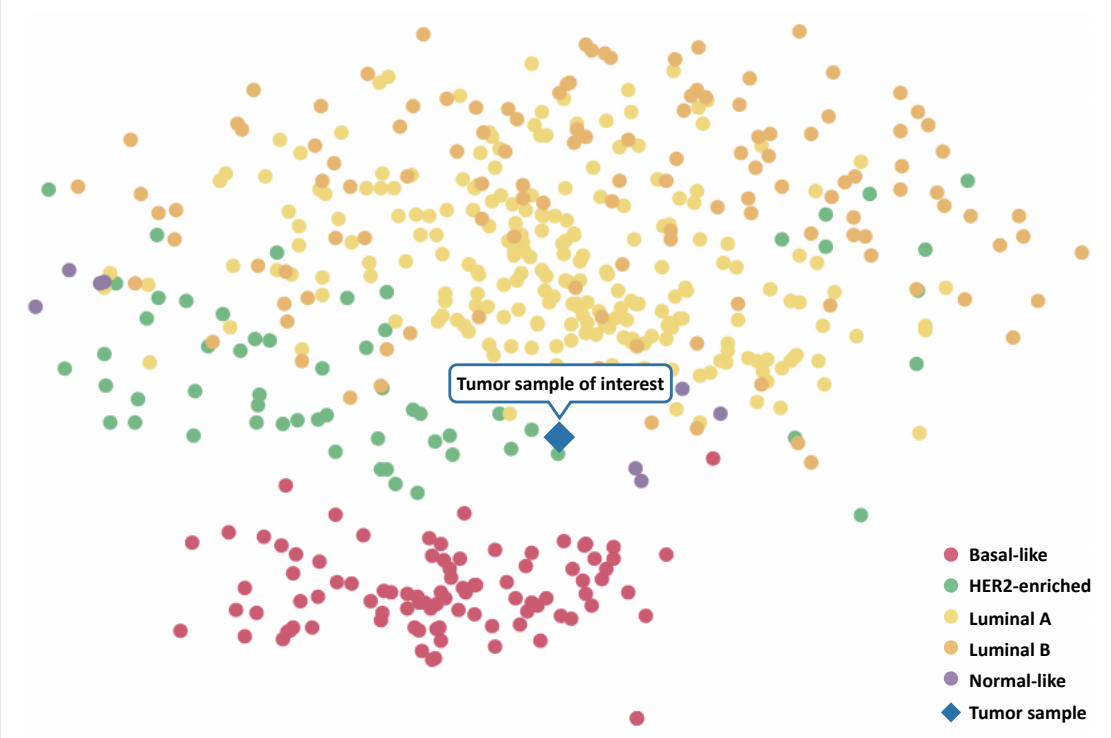


Figure 6.5 **T-SNE clustering in ClinOmicsTrail^{bc}**. An uploaded tumor gene expression sample (`TCGA-AN-A0XN`, **Section 6.5.1**) is clustered along with primary breast tumor samples from TCGA. The molecular subtypes of the TCGA samples are color-coded as indicated by the legend in the lower right corner. The tumor sample under investigation is indicated by the blue diamond-shaped symbol.

6.4.3 Decision support functionality

Based on the analyses described in **Section 6.4.2**, key tumor characteristics as well as the suitability of various types of drugs, both on-label and off-label, can be investigated (cf. **Figure 6.3**, fourth column). As a starting point, ClinOmicsTrail^{bc} provides a characterization of the given tumor regarding its specific genomic and transcriptomic alterations and their impact on affected signaling pathways (**Section 6.4.3.1**). The tumor under consideration is also analyzed with respect to its similarity to other tumors, allowing to classify the tumor subtype not just on the provided receptor status, but also based on its transcriptomic profile. ClinOmicsTrail^{bc} assesses a set of standard-of-care breast cancer drugs with respect to a variety of relevant factors, such as the status of the molecular drug targets, drug-processing enzymes, transporters, and involved pathways (**Section 6.4.3.2**). Also, putative candidates for drug repositioning are indicated and can be further investigated. Additionally, the potential suitability of immunotherapies is determined with respect to neoepitope vaccines and checkpoint inhibitors (**Section 6.4.3.3**). Finally, ClinOmicsTrail^{bc} performs a first assessment of the eligibility to participate in clinical trials (**Section 6.4.3.4**).

6.4.3.1 Overview of specific tumor characteristics

In order for the user to obtain a comprehensive overview of a tumor's specific characteristics, ClinOmicsTrail^{bc} provides an interactive visualization of selected driver genes and signaling processes within the tumor in the form of a sunburst chart (cf. **Figure 6.6**). Relevant signaling pathways and representative genes are displayed in a circular manner. The innermost ring represents cancer-relevant pathways. Each segment (i.e., pathway) is colored by its inferred pathway activity, as described in **Section 6.4.2**. Clicking on a pathway of interest zooms into this pathway for a focused representation of the data.

Depending on the types of provided omics data, up to seven additional rings are displayed: in the most comprehensive case the rings indicate (from inside out) sample-specific measurements of (1) differential gene expression, (2) (differential) methylation scores, (3) copy number alterations, (4) genomic mutations, (5) the corresponding gene's name, (6) indicators on whether the gene is an oncogene or a tumor suppressor, as well as its (7) druggability status. Somatic mutations are color-coded by the predicted type of mutation (e.g., missense, frameshift, stop loss). Here, we use the definition of oncogenes and tumor suppressor genes as proposed by Sanchez-Vega *et al.* in a recent publication [633]: Oncogenes are defined as those genes for which activating mutations (or other upregulating alterations) lead to an activation of their associated pathways. Analogously, tumor suppressor genes are defined as those genes for which inhibiting mutations (or other downregulating alterations) contribute to the pathways' activations.

The sunburst chart is also connected to other third-party resources allowing for a detailed investigation of specific genes or pathways. For example, clicking on a gene's name opens details from NCBI Gene [270]. Additional details on the specific mutation, as well as various scores indicating its severity (SIFT [221] and PolyPhen [634]), and links to external databases (dbSNP [226] and COSMIC [612]) can be obtained by clicking on a mutation of interest. Also, known pharmacogenomic relationships for the contained mutation are displayed.

This interactive overview visualization is fully searchable, zoomable, and extendable. Searching for a gene or pathway of interest will highlight the respective section in the plot. If one or several gene(s) of interest are not yet contained in the sunburst chart, these can be interactively added to the visualization via entering the gene's name into the search field mentioned earlier. The respective entries will be appended to the chart in a user-defined category.

6.4.3.2 Assessment of targeted therapies

When faced with the decision of which breast cancer drug(s) to prescribe a patient, clinicians typically assess a variety of clinical markers such as hormone receptors or menopausal status. Based on the provided multi-omics tumor data, ClinOmicsTrail^{bc} provides additional insights by considering several classes of genes, proteins, and pathways that might promote or hinder the effectiveness of a drug. For a set of 17 FDA-approved, standard-of-care breast cancer drugs (cf. **Figure 6.7**), ClinOmicsTrail^{bc} assesses the genomic and transcriptomic status of respective molecular drug targets, drug-processing enzymes, resistance-promoting factors, and associated pathways. Relevant factors to consider were obtained from DrugBank [427], the Therapeutic Target Database [635], and the literature. Since the respective categories reflect different mechanisms that might (de)sensitize a tumor with regard to the considered drug, different clinical, genomic, and transcriptomic traits have to be considered in each case. For

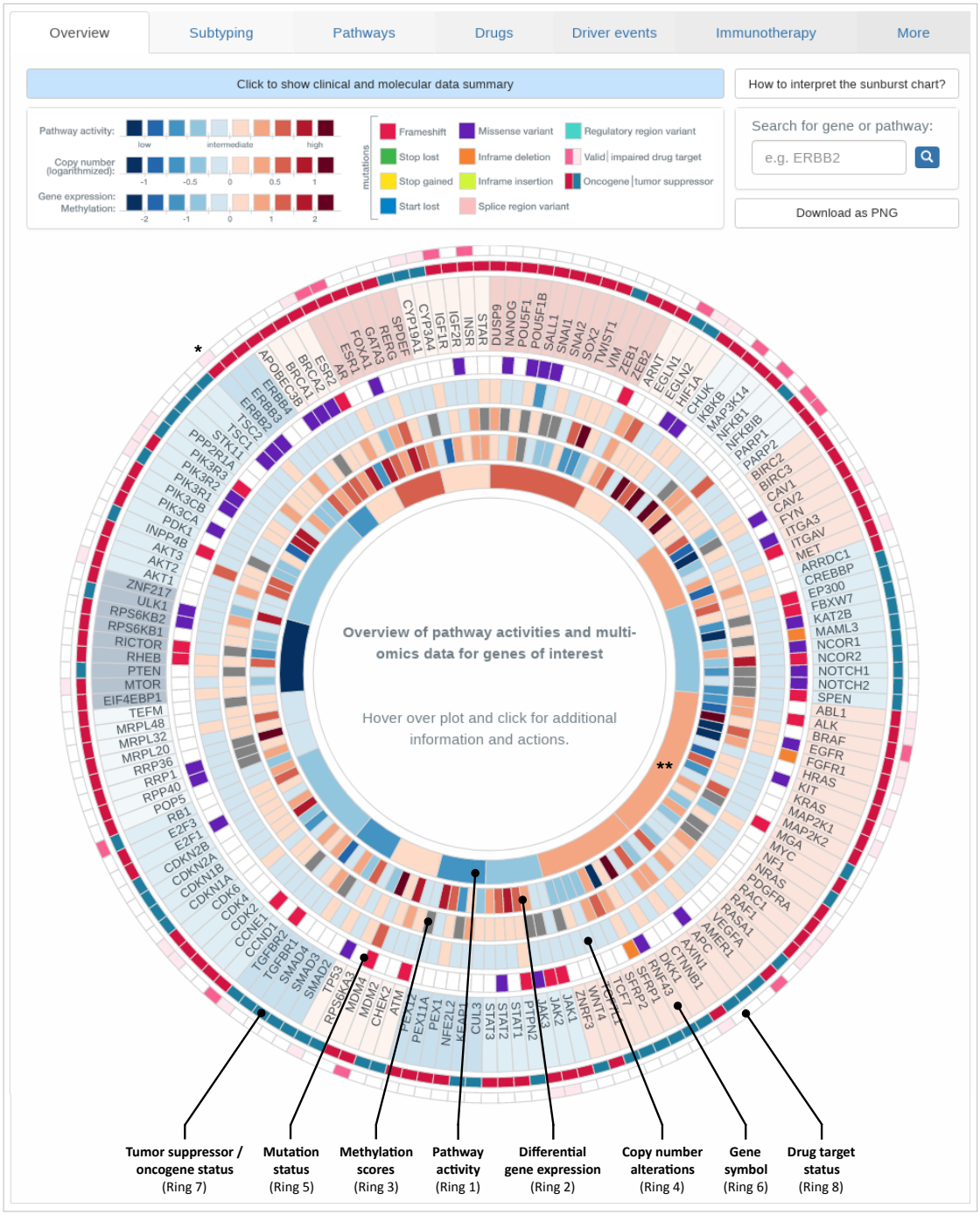


Figure 6.6 Overview of tumor characteristics in ClinOmicsTrail^{bc}. Breast cancer-relevant driver genes and pathways are displayed in a circular manner. Genes are grouped according to the pathways they are most characteristic for. The plot is organized in rings, where the innermost ring displays pathway activities, the second 'inner' ring corresponds to gene expression. Depending on the data provided by the user, information on copy number alterations and mutations are shown in the third and fourth ring, respectively. Gene names are displayed in the next ring. The second most outer ring indicates whether the gene acts as an oncogene or a tumor suppressor gene for activating the corresponding pathway. The outermost ring contains indicators on whether or not the gene is a known drug target. Visualization for sample TCGA-BH-A0DT (cf. **Section 6.5.2**). * Entry for HER2/neu (ERBB2), ** MAPK signaling pathway as referred to in **Section 6.5.2**.

many breast cancer drugs, there are known predictive biomarkers that inform the treatment decision-making process. Aromatase inhibitors, for example, are typically only administered to postmenopausal women with positive hormone receptor status [636], while amplification of the HER2/neu favors treatment with trastuzumab [637]. In ClinOmicsTrail^{bc}, predictive biomarkers like ER status, PR status, or HER2/neu amplification status are assessed first and foremost on the clinical data provided by the user. However, these clinical indications are also compared to gene expression and/or copy number data to spot potential inconsistencies in the data, i.e., if one or both molecular data sets disagree with the provided tumor receptor status.

Another important set of factors for drug efficacy are the molecular drug target(s) of a compound of interest. Here, it is favorable if the drug target is highly expressed in the tumor. ClinOmicsTrail^{bc} also investigates whether the drug target contains a somatic mutation and, if this is the case, assesses the mutation's severity to determine if the target might have attained a resistance mutation (e.g., compromised binding affinity of a drug or if the corresponding protein has lost its function). In this case, the targeting of this protein with a drug is likely to be ineffective.

When investigating putative drug efficacy, pharmacokinetic mechanisms also have to be considered. In this regard, ClinOmicsTrail^{bc} assesses drug-metabolizing enzymes as well as efflux transporters. Many drugs require activation by drug-metabolizing enzymes like cytochromes in the liver [638]. Alterations and especially germline mutations in this gene family are a major resource of variability in treatment response [639]. ClinOmicsTrail^{bc} uses the (germline) mutation status of the respective enzymes to determine whether or not drugs can be metabolized to their active forms. Also, efflux transporters are a potent cause of variability in drug response and even drug resistance [640]. When highly active, they carry the compound out of the cell, thus decreasing its intracellular concentration and hence its efficacy. Here, ClinOmicsTrail^{bc} especially focuses on gene expression data to detect increased transporter activity, but it also takes somatic mutations into account. However, a high transporter activity is not always associated with reduced drug efficacy. Some drugs (e.g., tamoxifen or lapatinib) can inhibit certain transporters (e.g., ABCB1) and hence improve the tumor's response to other drugs that are affected by high efflux transporter activities [641, 642].

As a final class of modulators of drug response, we also consider whole signaling pathways. Here, pathways directly targeted by the compound under consideration should show strong activities in the tumor. By this, we ensure that the drug actually tackles a disease-driving mechanism. ClinOmicsTrail^{bc} assesses the activity of a pathway based on a set of characteristic pathway-associated genes via their gene expression scores, see **Section 6.4.2**.

A summary of the rule-based system for drug evaluation is provided in **Section A.8.2**.

Besides the assessment of on-label drugs, ClinOmicsTrail^{bc} also investigates a set of 23 'driver targeting drugs'. These are drugs that require the presence or absence of pathological markers, mutations, or other genomic alterations in the *Driver targeting drugs* tab of the *Driver events* view. These drugs are approved for various cancer types, however, not necessarily for breast cancer. Still, they could be considered for off-label use. In order to determine whether or not a patient could be stratified for the administration of the respective drugs, ClinOmicsTrail^{bc} evaluates the genomic alterations in the tumor with respect to specific point mutations, copy number alterations, transcriptomic deregulation, and hormone receptor status.

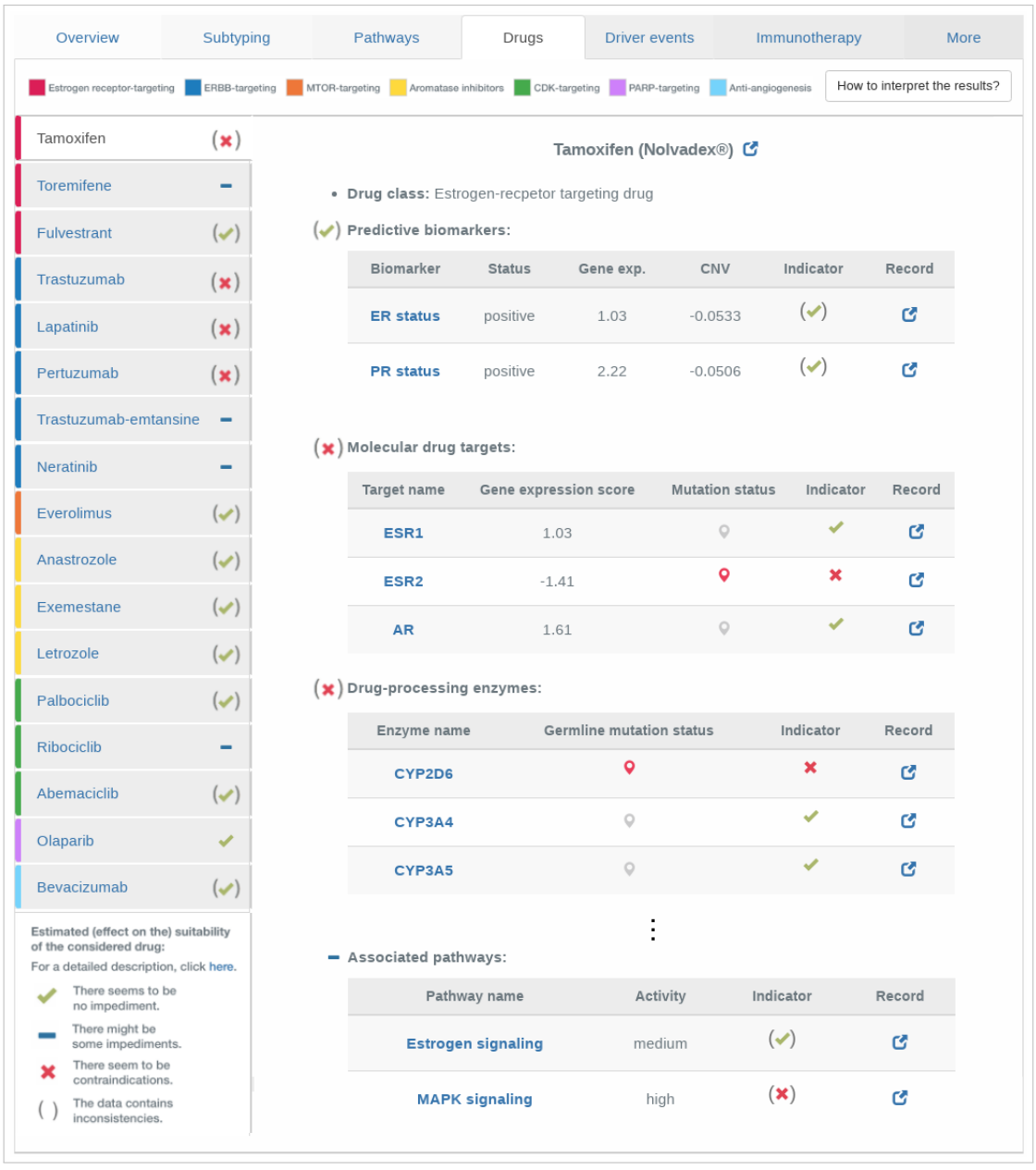


Figure 6.7 **Assessment of standard-of-care drugs in ClinOmicsTrail^{bc}.** For a set of 17 standard-of-care breast cancer drugs (left panel), various factors increasing or decreasing the efficacy of a drug are assessed. Clinical, genetic, and molecular characteristics are listed with an indicator sign on whether they might decrease efficacy or even cause resistance to the treatment with the drug under consideration. All genes and pathways are linked to third-party resources, where additional details can be found. Each entry also contains the link to a record or publication that describes the role of the corresponding gene with respect to the drug of interest. Here, the results for the exemplary sample TCGA-BH-A0DT are displayed (cf. **Section 6.5.2**).

6.4.3.3 Immunotherapy assessment

One of the central characteristics of tumors is their ability to evade detection and destruction by the immune system [49]. One such mechanism is the activation of immune-inhibitory pathways and immune system-suppressing checkpoints like CTLA-4 or PD-1 [82]. These checkpoints typically balance the immune system's activation and tolerance. This balance is often impaired in cancer and can lead to immune evasion of even heavily altered cells that present many tumor antigens. Typically, recognition of tumor antigens presented on HLA molecules on the surface of tumor cells by effector T-cells leads to the destruction of the tumor cells. However, in the presence of an expressed programmed death ligand 1 or 2 (PD-L1/2), this ligand may bind to the PD-1 receptor on the T-cell, inducing an inhibitory signal and preventing the tumor cell from being destroyed. In such cases, inhibiting either the receptor or the ligand (e.g., by pembrolizumab binding PD-1 [83]) might restore the immune system's ability to kill tumor cells. Several studies across cancer types showed that the effectiveness of such checkpoint inhibitors, amongst others, correlates with the number of mutations present in a tumor [85, 86]. The more mutations are contained in a cell, the more likely it is that non-synonymous mutations occur in human leucocyte antigen (HLA) epitopes that - when presented on the cell surface - may be recognized by cytotoxic T-cells, which in turn induce cell killing [643]. Tumors with a high mutational load have the potential to respond well to immune system inducing therapies like antigen vaccination or adoptive cell therapy [644, 645]. The total number of somatic mutations per coding region of the genome is defined as Tumor Mutational Burden (TMB) [86]. ClinOmicsTrail^{bc} computes the TMB as the number of somatic mutations per megabase of exon. Moreover, ClinOmicsTrail^{bc} displays the TMB of tumor samples of interest in comparison to the TCGA breast cancer cohort (cf. **Figure 6.8**) for a relative assessment of severity.

As high mutation rates are oftentimes caused by deficiencies in the DNA repair machinery, ClinOmicsTrail^{bc} also assesses the status of a variety of repair genes, curated by MD Anderson Cancer Center [646]. Additionally, ClinOmicsTrail^{bc} provides information on the gene expression and copy number status of HLAs and the checkpoint ligands.

Cancer vaccines: Besides checkpoint blockade, personalized cancer vaccines are another promising approach to cancer immunotherapy [88, 89]. Cancer vaccines target overexpressed or altered proteins and HLA presented peptide sequences (neoepitopes) that resulted from somatic mutations uniquely characterizing the patient's tumor. They are used to prime T-cells to recognize these characterizing antigens and destroy the presenting tumor cells. As the neoepitopes are dependent on both the patient's tumor mutations and HLA genotype, cancer vaccines have to be individually designed. Thus, ClinOmicsTrail^{bc} offers functionalities to predict potential neoepitope vaccine targets based on the identified somatic mutations and HLA genotype of a patient using the immunoinformatic toolbox ImmunoNodes [647]. ImmunoNodes provides various classes of epitope prediction methods to compute (neo-)epitopes and to assess their affinity to the patient's HLA genotype. **Section A.8.4** contains additional details on the 13 provided methods. The identified epitopes can serve as a basis for the synthesis of a personalized cancer vaccine, which can be combined with checkpoint inhibitors to potentially boost the effectiveness of the vaccine [90].

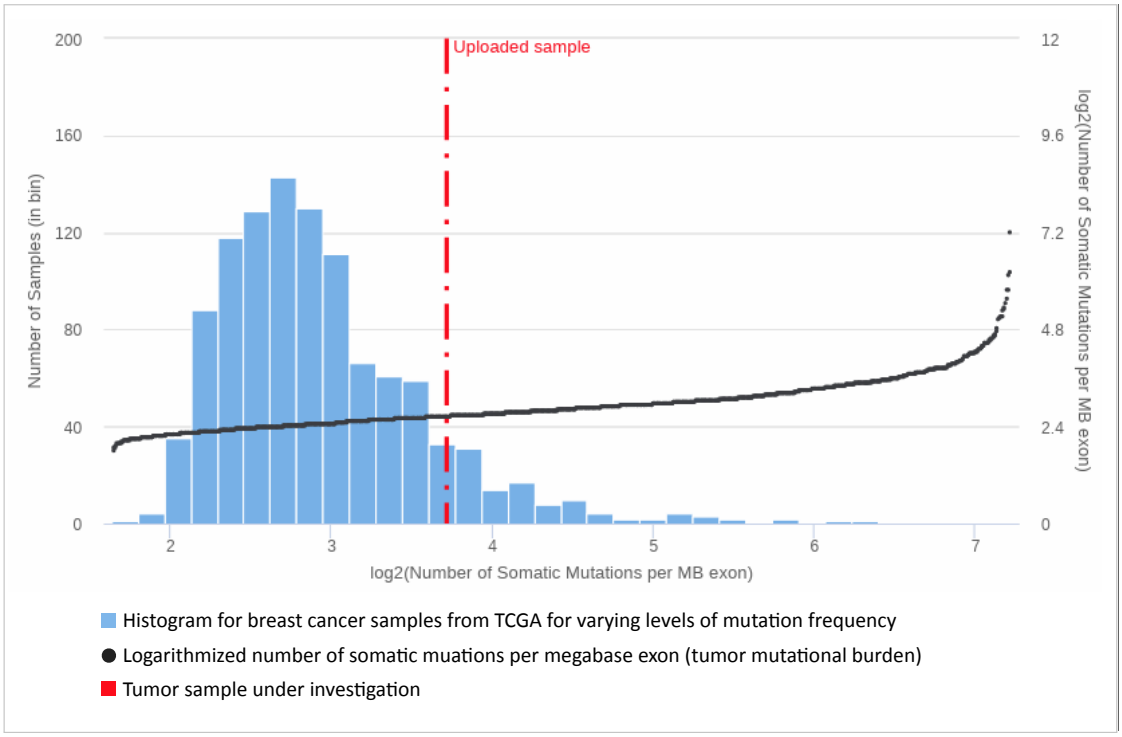


Figure 6.8 **Tumor mutational burden in ClinOmicsTrail^{bc}**. Visualization of the tumor mutational burden for a sample of interest (`TCGA-A2-A0T2`, in red, cf. **Section 6.5.3**) in comparison to the TCGA breast cancer cohort. The blue bars indicate the number of TCGA samples per interval of mutation frequencies (left y-axis). The TCGA samples are sorted by increasing mutation load. The black dots depict the logarithmized number of somatic mutations per megabase exon (right y-axis).

6.4.3.4 Clinical trial matching

In cases where standard-of-care treatment solutions are not applicable due to resistance mutations or other hindering factors, it might be of interest to examine potential clinical trials the patient is eligible to enroll in. To this end, ClinOmicsTrail^{bc} links to phase II, III, and IV clinical trials registered in *ClinicalTrials.gov* and the *European Union Clinical Trials Register*, which are recruiting in a large variety of countries. Additionally, ClinOmicsTrail^{bc} makes a first assessment of the eligibility for various classes of clinical trials listed on *BreastCancerTrials.org*. This stratification considers tumor characteristics like the BRCA1/2 mutation status and the tumor grade, as well as different treatment types, including hormone therapy, PARP inhibitors, targeted therapy, and immunotherapy.

6.4.3.5 Investigation of deregulated drug targets and signaling processes

For an even more in-depth investigation of deregulated drug targets and altered signaling processes in the tumor, ClinOmicsTrail^{bc} is natively integrated with its sister projects DrugTargetInspector (cf. **Chapter 5**) and GeneTrail2 (cf. **Section 4.2**). DrugTargetInspector is a web service for the interactive investigation of drug targets and dysregulated signaling pathways. GeneTrail2 is a web-interface providing access to different tools for the statistical analysis of molecular signatures with a focus on enrichment analyses. It offers multiple statistical tests, as well as a comprehensive collection of biological gene sets to test for. Once omics data sets are uploaded to ClinOmicsTrail^{bc}, the functionality of both tools is readily available.

6.5 Case studies

In order to show ClinOmicsTrail^{bc}'s rich functionality and its potential to support the clinical cancer treatment decision-making process, we performed three case studies in which ClinOmicsTrail^{bc} (i) guides the treatment selection process by identifying pathway activity patterns driving the tumor under investigation (**Section 6.5.1**), (ii) assesses a set of drugs approved for breast cancer treatment by an in-depth investigation of modulators of treatment success (**Section 6.5.2**), and (iii) highlights immunotherapy as a potential treatment option in cases with high tumor mutational burden (**Section 6.5.3**). Additional supplementary figures for the three case studies are provided in **Section A.8.5**.

6.5.1 Pathway activity patterns guiding treatment selection

For the identification of a tumor's intrinsic subtype, various systems using multigene signatures have been established, most famously the PAM50 classifier [599] (cf. **Section 6.3**). However, the consideration of a broader spectrum of gene expression might draw a more differentiated picture of a tumor's characteristics. For instance, the consideration of pathway activity patterns might be useful to guide the treatment decision-making process.
For illustratory purposes, we considered the tumor sample of a 68-year-old woman (`TCGA-AN-A0XN`) with a stage III breast cancer (ER negative, PR positive, HER2 negative) that was classified as luminal A subtype by PAM50. ClinOmicsTrail^{bc} provides various analyses for the in-depth investigation of a tumor's molecular characteristics and subtype. For an assessment

of molecular similarities to other breast tumors, the *Subtyping* tab contains a clustering-based visualization that embeds the tumor sample under investigation into a 'neighborhood' of TCGA BRCA samples with similar gene expression signatures (cf. **Section 6.4.3**). For our sample of interest, we can observe that, although it was predicted to be of the luminal subtype, it also has similarities to the HER2-enriched and basal-like subtypes (cf. **Figure 6.5**). For a more granular investigation of the tumor's characteristics, ClinOmicsTrail^{bc} computes the pathway activities of a set of 20 core breast cancer-relevant pathways (cf. **Section 6.4.2**) that can exhibit subtype-specific patterns. The pathway activities depicted in the *Pathways* view show that in our sample of interest – in contrast to the majority of breast tumors – the PI3K-Akt signaling pathway seems to be inactive, possibly due to a most likely damaging frameshift mutation in PIK3CA, one of the pathway's core activating elements. On the contrary, several pathways related to stem cell characteristics (e.g., Focal adhesion and HIF-1 signaling) seem to be strongly upregulated (cf. **Figure 6.4**, in blue). The activation of these pathways has already been shown to be characteristic of basal-like tumors [648, 649]. These findings are supported by the fact that the sample under investigation shows a pathway activity pattern that is very similar to the one of the triple-negative, basal-like TCGA sample `TCGA-BH-A18G`, see **Figure 6.4**. Furthermore, the ten cell lines most similar to our tumor sample of interest are all triple-negative (cf. **Section A.16**). The investigation of these pathway activity patterns might also reveal targets for possible therapeutic intervention. In the considered sample, the MAPK signaling pathway is strongly activated, most likely due to the upregulation of key pathway components like BRAF, KRAS, and NRAS. Therapeutic intervention in this pathway might hence be an option. This finding is supported by the fact that the two cell lines most similar to the sample under investigation regarding pathway activity patterns (`SUM149PT` and `185B5`) are known to be sensitive against an ERK inhibitor. Another potential option suggested by ClinOmicsTrail^{bc} might be a treatment with bevacizumab as the vascular epithelial growth factor (VEGFA), its molecular target [72], is very strongly upregulated (z-score=3.63).

6.5.2 Assessment of standard-of-care breast cancer drugs

The selective estrogen receptor modifier tamoxifen is one of the oldest and most commonly prescribed breast cancer drugs [650]. Its clinical benefit for the treatment of estrogen receptor-positive breast cancer is well established by significantly reducing mortality rates and recurrence [651]. Still, more than 30% of patients with adjuvant tamoxifen treatment relapse or die [652]. This is likely due to *de novo* or acquired tamoxifen resistance that can be mediated by a variety of genetic and molecular factors, as well as altered signaling pathways. These resistance-promoting factors include lowered expression or mutation of molecular drug targets, the impairment of involved ADME genes, alterations in co-regulatory proteins, and deregulated signaling cascades [653]. In order to exemplify ClinOmicsTrail^{bc}'s thorough assessment of these factors, we consider the tumor sample of a 41-year-old woman with a stage II, hormone receptor-positive, HER2-negative breast tumor (`TCGA-BH-A0DT`). As the considered sample is hormone receptor-positive and HER2-negative, tamoxifen might be considered as the treatment of choice. In order to obtain a more comprehensive picture, ClinOmicsTrail^{bc} provides a *Drugs* view that contains information on the status of several biomarkers potentially relevant to estimate treatment success like molecular targets, drug-processing enzymes, and transporters (cf. **Figure 6.7**). Tamoxifen targets, besides the androgen receptor (AR), the estrogen receptors 1 and 2 (ESR1/ERα,

ESR2/ERβ), a family of transcription factors that are activated by estrogens, mediating the activation of a variety of growth-promoting processes [654]. In our sample under investigation, ESR2 contains a frameshift variant that is likely to affect the protein's structure severely and hence might drastically reduce tamoxifen's affinity to its target ESR2. Besides the actual drug targets, different ADME genes, including drug-processing enzymes and transporters, play essential roles in the effectiveness of a tamoxifen-based therapy. A central element in this context is the Cytochrome P450 family member CYP2D6 that is known to metabolize up to 25% of commonly prescribed drugs, including the prodrug tamoxifen that needs to be metabolized into its active form endoxifen [621]. For our given sample, ClinOmicsTrail^{bc} identified a frameshift variant in CYP2D6 that most likely generates a poor metabolizer phenotype and hence contributes to a potential resistance against tamoxifen. Another group of relevant factors are co-regulators of ER-mediated transcription. One such regulatory element is the cytosine deaminase APOBEC3B, which typically deaminates cytosine to uracil in ER enhancer regions, thereby activating base excision repair pathways, which in turn promote chromatin remodeling that eventually help to initiate the expression of ER target genes [655]. Higher levels of APOBEC3B expression have been associated with poor clinical outcome of tamoxifen treatment in ER-positive breast cancer [427]. In our sample of interest, we can observe an increased level of APOBEC3B expression, serving as a further indication in disfavor of tamoxifen. Finally, the considered sample shows increased levels of MAPK signaling pathway activity and an upregulation of HER2/neu (ERBB2) (cf. **Figure 6.6**), which might contribute to resistance against endocrine therapy via the ligand-independent activation of ER through ERK [653, 656]. To summarize, although the clinical data for the considered sample might point towards a treatment with tamoxifen, a broad investigation of molecular determinants of treatment success could highlight several factors that might render tamoxifen ineffective in the considered case. Besides selective estrogen receptor modulators like tamoxifen, ClinOmicsTrail^{bc} also provides an in-depth assessment of a variety of other drug classes relevant for breast cancer treatment (cf. **Figure 6.7**). Although we could observe an upregulation of ERBB2 expression in our sample under investigation, trastuzumab and other ERBB-targeting drugs might be impeded by a missense variant in ERBB2. This mutation could reduce the efficacy of this class of drugs, despite the fact that it has been classified to only have a moderate impact on the protein's structure. As indicated by ClinOmicsTrail^{bc}, a putative treatment option for our investigated sample might be the use of aromatase inhibitors like anastrozole, exemestane, or letrozole. As aromatase inhibitors are typically prescribed to postmenopausal women [657], a successful treatment will require additional ovary suppression, which has been shown to significantly improve response rates in premenopausal women in the SOFT trial [658]. Also, since our sample of investigation contains BRCA1/2 germline mutations and the poly ADP ribosyl transferase PARP1 is strongly upregulated (z-score=4.99), PARP inhibitor treatment is suggested as a potential option by ClinOmicsTrail^{bc}.

6.5.3 Immunotherapy assessment

One of the latest areas of innovation in cancer treatment is immunotherapy, which aims at (re-)enabling the immune system to recognize and destroy cancerous cells [659]. While being most established for the treatment of melanoma [660] and lung cancer [661], immunotherapeutic approaches like checkpoint blockade or cancer vaccines are of general

nature and not necessarily tumor type-specific. Over the last years, more and more immunotherapies have been approved for a variety of cancer types [662–664]. Immunotherapies are also emerging as a valuable component of treatment regimens in breast cancer. Many studies show that tumors with a high Tumor Mutational Burden (TMB) are likely to respond well to an immune system-promoting therapy [644, 645]. In order to evaluate the immunotherapeutic potential of a breast tumor sample under investigation, ClinOmicsTrail^{bc} assesses the sample's TMB in relation to the TCGA BRCA cohort. As a reference, TMBs for 1,044 breast tumor samples were computed and plotted in a histogram indicating the respective fractions of samples containing a certain number of mutations (cf. **Figure 6.8**). In this reference group, the mutational burden ranges between 3.49 and 148.57 somatic mutations per megabase exon (for variants called using MuTect2 [210]). In samples with high TMB, the increased mutation rates are likely to be fostered by deficiencies in the DNA repair machinery [6]. To further underline the connection between TMB and an impaired DNA repair machinery, we performed the following analysis: from the aforementioned TCGA BRCA cohort, we selected the 100 samples with the lowest and highest TMB, respectively. For both groups, we extracted those genes that contained mutations with a high disruptive functional impact, e.g., via protein truncation, loss of function, or nonsense-mediated decay, as annotated by Ensembl's Variant Effect Predictor. These gene sets were then tested in an Over-Representation Analysis (cf. **Section 3.3.3.1**) for the enrichment of a set of 52 repair-related biological categories obtained from GO, KEGG, Reactome, and WikiPathways. For the 100 samples with low TMB, none of the considered categories were significantly enriched, whereas for the 100 samples with high TMB, 32 pathways showed a significant enrichment (cf. **Table A.13**).

For a convenient assessment of potential deficiencies in the DNA repair machinery, ClinOmicsTrail^{bc} assesses a variety of repair genes with respect to their mutation status and (epi-)genetic profile. This information can also be used to guide the immunotherapy treatment-selection process as tumors with loss of mismatch repair function are likely to avoid immune system-mediated destruction through the activation of immune checkpoints [665] and hence might become sensitive to checkpoint blockade.

As a concrete example demonstrating ClinOmicsTrail^{bc}'s capabilities for decision support in immunotherapy, we considered the tumor sample of a 66-year-old woman with a stage IV, triple-negative, metastatic breast cancer (`TCGA-A2-A0T2`). Based on the genomics and transcriptomics data of a tumor sample, ClinOmicsTrail^{bc} analyzes potential breast cancer driver genes and lists their genomic and transcriptomic features, including the effects of contained mutations in the *Driver events* view. Here, the analysis revealed severe and probably damaging mutations in TP53, RB1, and ATM. The transcription factor TP53 is an essential tumor suppressor that is commonly compromised in human cancers. The encoded protein is involved in a variety of cellular processes, including cell cycle arrest, senescence, apoptosis, and DNA repair [666]. Similarly, RB1 acts as a tumor suppressor by negatively regulating the cell cycle. It is also involved in stabilizing heterochromatin, thereby maintaining the overall chromatin structure [667]. Hence, alterations in RB1 can cause genomic instability, fostering the accumulation of mutations and providing an evolutionary advantage to the affected cancer cells [668]. Finally, ATM is an important cell cycle checkpoint kinase that regulates various tumor suppressors like TP53 or BRCA1, acting as key regulators governing genome stability and response to DNA damage [669].

Besides known driver mutations, ClinOmicsTrail^{bc} also checks the mutation status of a variety of genes involved in the DNA repair machinery and provides an overview of potentially impaired genes in the *Repair genes* tab of the *Immunotherapy* view. For our sample under investigation, these results revealed a variety of mostly damaging mutations in various components of DNA repair (cf. **Table 6.2**). The impairment of repair mechanisms was also reflected in a rather high TMB of 13.18 somatic mutations per megabase exon that is illustrated in the *Mutational burden* tab. Taken together, the mutational burden in combination with the likely impairment of repair mechanisms might render checkpoint blockade (potentially in combination with DNA-damaging agents [670] or neoepitope vaccination [671]) an effective treatment strategy in this case [82, 672]. For further assistance in the selection of a suitable checkpoint inhibitor, ClinOmicsTrail^{bc} provides an overview of the genomic and transcriptomic features of various druggable checkpoint genes.

Cancer vaccines: With respect to personalized cancer vaccines, ClinOmicsTrail^{bc} identifies tumor-specific neoepitopes that can serve as a basis for vaccine development. Details on the neoepitope prediction will be described in the following section.

Gene	Consequence	SIFT	PolyPhen	Reference
ATM	missense variant	deleterious	probably damaging	[669]
APTX	missense variant	deleterious	probably damaging	[673]
CCNH	missense variant	deleterious	probably damaging	[674]
FANCC	missense variant	deleterious	probably damaging	[675]
FANCF	missense variant	deleterious	benign	[676]
MSH5	missense variant	tolerated	benign	[677]
POLG	missense variant	deleterious	possibly damaging	[678]
RAD54B	missense variant	deleterious	possibly damaging	[679]
XAB2	missense variant	deleterious	probably damaging	[680]

Table 6.2 **Mutated repair genes in sample `TCGA-A2-A0T2`**. The first column contains the HUGO gene symbol of the respective gene and the second column the effect of the mutation on the protein-coding sequence. Columns 3 and 4 list predictions on the effect of a mutation on the protein function made by SIFT and PolyPhen. The last column contains literature references explaining the role of the corresponding gene in the repair machinery.

The genomic regions encoding for human leucocyte antigen (HLA) proteins are very polymorphic, hence the HLA genotype strongly varies between patients. In order to identify those tumor mutations that are likely to be able to induce an immune response, the patient's HLA genotype and its tumor's somatic mutations have to be taken into account. The HLA genotype can be experimentally determined (e.g., using sequence-specific oligonucleotide probe hybridization [681] or serological typing techniques [682]), but also predicted from sequencing data. For illustratory purposes, we here applied the HLA genotyping algorithm OptiType [683] to the raw tumor sequencing data for our sample of interest. OptiType predicted the sample to be of the following genotype: A*02:01, A*24:02, B*15:17, B*40:01, C*07:01, C*03:04.
Within ClinOmicsTrail^{bc}, we selected the option *Consider only significantly upregulated proteins* in the *Cancer vaccines* tab of the *Immunotherapy* view and performed a neoepitope prediction for

peptides of length 9 using the artificial neural network-based regression method NetMHC [684]. **Table A.15** lists all neoepitopes that were predicted to bind to at least on the HLAs.

6.6 Discussion

We presented ClinOmicsTrail^{bc}, a powerful visual analytics tool for breast cancer treatment stratification. Our tool supports precision medicine (i) by assessing and prioritizing standard-of-care breast cancer drugs, (ii) by suggesting drugs for off-label use, (iii) by evaluating the potential of different types of immunotherapy including checkpoint inhibitors and personalized cancer vaccines, and (iv) by assessing a patient's eligibility to enroll in clinical trials. To this end, ClinOmicsTrail^{bc} performs a multitude of analyses on a tumor's clinical markers, (epi-)genomic, and transcriptomic alterations to systematically characterize the tumor. This characterization is based on the tumor's main driver mutations, its mutational burden, the activity patterns of cancer-relevant pathways, as well as the status of drug-specific predictive biomarkers, molecular drug targets, and involved ADME genes.

In order to optimally support clinicians, conciseness and interpretability of the results are essential. For that purpose, ClinOmicsTrail^{bc} summarizes key tumor characteristics and additional information on drug-specific biomarkers and modulators of treatment response in a few comprehensive, yet easily interpretable visualizations, providing clinicians with the most relevant information at the point of care. Furthermore, we provide extensive documentation of the web service, ranging from standalone tutorials over additional help and information along the data upload and analysis steps to interactive explanations of the provided results. Albeit ClinOmicsTrail^{bc} is optimized for the analysis of breast cancer data sets, the underlying analysis methods and visualization techniques offered by our web service can also be used for the genetic and molecular characterization of other tumor types by mainly exchanging the tumor-specific underlying databases. We plan to provide adapted versions for other tumor types in the near future. The breadth and depth of analyzes and visualizations offered by ClinOmicsTrail^{bc} make it – although being in a proof-of-principle stage - a promising addition to existing clinical decision support machineries. The three presented case studies convey a first impression on the capabilities of the tool, however, can only partly illustrate its full potential. From a clinical perspective, ClinOmicsTrail^{bc} is a comprehensive tool suite that will be further validated regarding its benefits in the preparation and conduction of molecular tumor board meetings. In summary, ClinOmicsTrail^{bc} is a powerful integrated visual analytics tool for breast cancer research in general and therapy stratification in particular, assisting oncologists to find the best possible treatment options in a deeply personalized way.

7 Perspectives

In the history of biomedical research and patient treatment, diagnostic capabilities have evolved from the anatomical, over the cellular, to the molecular level [685]. Nowadays, high-throughput experimental techniques allow for the determination of comprehensive profiles of a broad range of biological data (cf. **Section 3.1**). While the genetic and molecular elucidation of healthy and aberrant mechanisms has advanced the general knowledge of the causes of numerous diseases [686], the translational success has been limited for complex diseases like cancer [687] (cf. **Section 2.2**). The concept of personalized medicine aims at improving (cancer) treatment via the tailoring of treatment options to the specific genetic and molecular characteristics of a disease. Nowadays, personalized medicine comes in various forms, including companion diagnostics and increasingly also gene panel testing (cf. **Section 2.3**). While these diagnostic tests have the potential to achieve considerable successes [110, 564, 688, 689], there are still major limitations mainly due to the fact these tests cannot capture the high complexity of aberrant processes in tumors [516, 690] (cf. **Section 2.1**).

In order to obtain a more comprehensive view on aberrant disease processes and to employ this information for the identification of biomarkers and for clinical treatment decision-making, powerful tools and methods are required. They should be able (i) to integrate a variety of heterogeneous, noisy, and high-dimensional data sets, (ii) to extract *a priori* biological and medical knowledge from relevant databases, and (iii) to create explorative tools for intuitive and concise visualization of the results [691]. Following these principles, the goal of this thesis was to develop methods and tools to perform multi-omics integrative analyses for decision support systems in personalized cancer treatment.

In the following sections, we will review the presented work (**Section 7.1**) and highlight challenges and opportunities for further developments (**Section 7.2**).

7.1 Summary and discussion

In this thesis, we presented a comprehensive suite of tools and methods for translational research and clinical decision support in oncology. The provided methods and tools offer rich functionality for the genetic and molecular characterization of tumors with an emphasis on deregulated biological processes and the identification of disease-driving regulatory key players. The identified tumor characteristics are then combined with *a priori* knowledge from clinical practice guidelines and relevant medical, pharmacological, and biological databases for a personalized assessment of various types of treatment options, including standard-of-care targeted drugs, candidates for drug repositioning, and immunotherapy.

In order to ensure interoperability and ease of use, our tools are based on a common underlying framework, called Graviton (cf. **Section 4.1**). In Graviton, the analysis tools are deployed as web services, which has the advantage of easy access for users and a single instance, which allows for centralized maintenance. On the downside, the upload of tremendously large data sets containing potentially sensitive data is not always feasible.

The analyses provided by the tools in our tool suite are based on the integrated analysis of different types of omics data sets with *a priori* knowledge from various databases. The quality of the analysis results crucially depends on the quality and comprehensiveness of these resources. Besides potential technical noise and inaccuracies that can occur when capturing omics data sets (cf. **Section 3.1**), a 'bias' can also be introduced by the sample itself, for example when biopsies only contain low percentages of tumor tissue or have captured a sub-population of tumor cells that are not representative (anymore) for the evolving tumor. Hence, in an ideal world, multiple samples from different locations of the tumor would be the optimal basis for any analysis. Besides the biological input data, also the content, comprehensiveness, and quality of databases containing the *a priori* knowledge used in our analyses are of critical importance (cf. **Section 3.2**). Here, a major limitation is given by the current knowledge and representation of regulatory and signaling pathways and their respective topologies [363, 692].

When investigating aberrant processes in cancer, we work with data of various nature (i.e., continuous, discrete, structured) and data of various quality (e.g., noise, non-biological variance). In order to identify differentially expressed (or methylated) genes (or proteins, miRNAs), there are several scoring methods (cf. **Section 3.3.2**), which we also implemented in our tools (cf. **Section 4.1.3**). Each of these methodologies elucidates different aspects of the data and hence should be carefully chosen. For example, when using z-scores, one has to consider that small changes in genes with low variances are 'inflated' and cannot necessarily be distinguished from more substantial changes in genes with higher variances. In contrast to z-scores, fold changes provide a more direct comparison between two entities, however neglecting the natural variability of the gene's expression. Besides the selection of a scoring method, also which samples are compared to each other is a point to consider. While the comparison of a tumor sample to a healthy control reveals which pathogenic processes are up- or downregulated, the comparison of tumor vs. tumor provides a more fine-grained view on the specific characteristics of the sample of interest in comparison to other tumors of the same type or subtype. Hence, the consideration of both types of comparisons would be desirable.

However, healthy reference samples are not always available, especially when only a single biopsy was taken from the tumor. As a remedy, it would be desirable to have a tissue-specific collection of reference samples, ideally all measured on the same experimental platform to minimize batch effects.

Genetic and molecular biomarkers are valuable tools to personalize diagnosis, treatment decision-making, and disease monitoring (cf. **Section 2.3**). However, the identification of robust biomarkers from extremely high-dimensional (multi-)omics data sets is a challenging task [402–404, 693], which is the reason that only few such biomarkers could be translated into clinical practice [601]. In order to support the identification of meaningful biomarkers for biomedical research and clinical decision support, we focused on the development of tools and methods that aim at elucidating causal dependencies instead of yielding 'black box' predictions that are solely based on mathematical criteria.

As a first tool in this context, we presented GeneTrail2 (cf. **Section 4.2**), which is, at the time of writing, one of the most comprehensive web services for enrichment analysis. Providing a great variety of functional annotations and pathway information, GeneTrail2 gives insights into aberrant biological processes and their functional dependencies. As an example of GeneTrail2's capabilities, we demonstrated how GeneTrail2 could support the identification of a molecular subtype of pancreatic cancer that is characterized by co-activation of the SUMO pathway and the oncogene MYC and which could be further investigated as a predictive biomarker for targeted therapy in pancreatic cancer (cf. **Section 4.2.3**).

When investigating mechanistic dependencies and key elements of pathogenic processes, the consideration of transcriptional regulators as the actual mediators of occurring aberrations in the cell is an important aspect. We presented RegulatorTrail (cf. **Section 4.3**), which provides four different analysis scenarios and a comprehensive set of analysis methods to identify deregulated regulatory processes. As one of the implemented methods, we proposed REGgulator-Gene Association Enrichment (REGGAE). We could show that REGGAE's unique approach of combining associations between regulators and their target genes with an enrichment approach outperforms competing methods in prioritizing the influence of transcriptional regulators (cf. **Section 4.3.3**). Moreover, we performed a case study on Wilms tumors in which we could identify the helix-loop-helix transcription factor TCF3 as a potential master regulator in the blastemal subtype, which could serve as the basis for the development of stratified treatment options (cf. **Section 4.3.4**).

While the presented tools can yield valuable hypotheses on potential disease mechanisms and corresponding biomarkers, it is clear that these hypotheses have to be further validated in wet-lab experiments and clinical studies.

Successfully validated biomarkers can then be used, for example, in the treatment stratification process. As current state of the art in personalized cancer treatment, genetic biomarkers in the form of specific mutations in selected genes are used to inform the choice of targeted treatment options (cf. **Section 2.3**). However, the vast majority of genetic aberrations contained in a tumor has not yet been assigned any clinical significance [694]. Hence, when reporting only variants annotated as (likely) pathogenic, many impactful aberrations might not be considered [695]. As a remedy, we additionally consider the predicted functional impact of mutations contained in the tumor under investigation, as they can provide valuable insights that otherwise might have been missed (cf. **Sections 5.3.2** and **6.4.1**). Moreover, for the identification of treatment-relevant tumor characteristics, we consider additional omics data types like transcriptomics, methylomics, or proteomics data to provide a multi-faceted picture of the tumor.

As our first tool with a specific focus on the assessment of targeted treatment options for drug repositioning and clinical decision support, we presented DrugTargetInspector (DTI, cf. **Chapter 5**). DTI provides information on deregulated drug targets, enriched biological pathways, and deregulated subnetworks. We demonstrated DTI's comprehensive functionality for the analysis of different types of omics data, as well as its ability to prioritize putative treatment options in several case studies (cf. **Section 5.3**).

With ClinOmicsTrail^{bc} (cf. **Chapter 6**), we developed a tool to dive even deeper into clinically relevant aspects of breast cancer. ClinOmicsTrail^{bc} provides a comprehensive assessment of standard-of-care targeted drugs, candidates for drug repositioning, and immunotherapeutic approaches. In three case studies, we demonstrated how ClinOmicsTrail^{bc} could facilitate

personalized treatment decisions in breast cancer based on actionable, evidence-based results (cf. **Sections 6.5.1** to **6.5.3**). In order to optimally support clinicians, conciseness and interpretability of the results are essential. To this end, ClinOmicsTrail^{bc} summarizes key tumor characteristics and details on drug-specific biomarkers in a comprehensive, yet easily interpretable way. Specifically, we predict the activity of tumor-specific signaling pathways by aggregating relevant genes based on a large body of prior biological and pharmacological knowledge. We believe that the consideration of tumor aberrations on a 'higher' pathway level is of great importance for the development of reliable and interpretable clinically relevant predictive models. In order to make the tool as concise and relevant as possible, we focused - for a first proof-of-principle version of the tool - on treatment-relevant aspects specific to breast cancer. However, the underlying analysis methods and visualization techniques offered by our web service can also be used for the genetic and molecular characterization of other tumor types by mainly exchanging the tumor-specific underlying databases. ClinOmicsTrail^{bc} is currently under investigation for the use in Molecular Tumor Board meetings at the university hospital in Tübingen.

While the assessment of deregulated signaling cascades and pathway activities is of great interest for basic cancer research and personalized treatment stratification, it is equally challenging. A major obstacle when trying to model pathway activities is the fact that there is no objectively measurable ground truth that could be used to optimize or validate the respective models. Hence, one has to rely on assumed proxies as the sensitivity to drugs targeting specific pathways or prior knowledge on the regulatory effect of individual genes. Moreover, the fact that biological pathways and signaling cascades form complex networks makes it even more difficult to link these observations to the activity of individual pathways. Another surrogate we currently still mainly use is mRNA expression instead of actual protein levels. Once experimental methods allow to routinely and comprehensively assess protein levels, the quality of models that aim at elucidating mechanistic dependencies in cancer is likely to improve.

To summarize, in this thesis, we have presented a comprehensive tool suite for cancer treatment decision support and translational research. The encompassed tools provide rich functionality for the genetic and molecular characterization of tumors with an emphasis on deregulated biological processes and the identification of disease-driving regulatory key players. In several case studies, we could show that the combination of the tumor characteristics identified by these tools in combination with *a priori* knowledge from clinical practice guidelines and relevant medical, pharmacological, and biological databases facilitates treatment decision support for various cancer types and several types of treatment options including standard-of-care targeted drugs, candidates for drug repositioning, and immunotherapy.

7.2 Outlook and conclusion

The tools and methods described in this thesis are a necessary and promising first step towards the multi-omics integrative analysis of tumors. However, the comprehensive implementation of precision oncology into routine clinical practice still faces many challenges.

One of the major challenges is that the breadth and depth of molecular and clinical data to be considered in personalized diagnosis, treatment, and monitoring is likely to keep growing. Although sequencing costs have decreased substantially during the last decades, to date genome sequencing is typically only considered for late-stage tumors and oftentimes only using

somewhat limited panel sequencing techniques. Prospectively, it can be expected that panel sequencing will be superseded by whole-exome or even whole-genome sequencing in the future, which will allow for the comprehensive analysis of tumor genomes without any limitation to predefined candidate genes.

A whole new level of complexity could additionally be introduced by the continuously developing molecular profiling techniques, which might make high-resolution techniques like single-cell sequencing feasible for clinical practice, thereby providing an even clearer picture of tumor subclones and their respective mutation patterns.

Besides the increasing resolution of molecular characterizations of tumors, also other data types, for example imaging data from radiology scans, need to be integrated and analyzed to obtain an even more complete perspective on tumors under investigation.

Another important aspect that increases the challenge of integrating and interpreting the data is that not just a single snapshot of clinical and molecular characteristics should be considered to describe complex diseases like cancer, but these data should instead be monitored longitudinally. Also, the patient's history and potential comorbidities need to be taken into account, as well as the effects of comedication or combination therapies.

This steadily growing amount of increasingly heterogeneous and complex data types calls for the continuous development and improvement of bioinformatics and statistical methods to integrate and analyze these data. An emerging concept whose adoption would support this challenging task is the concept of FAIR data, where FAIR stands for Findable, Accessible, Interoperable, and Reusable. While interoperability and reusability of data are typically limited by technical aspects like inconsistent nomenclatures across different databases or different experimental platforms and processing pipelines, the regulation of access to sensitive (patient) data accounting for data security and privacy according to the policies of the General Data Protection Regulation (GDPR) is a major challenge in itself.

In summary, there is clearly enormous potential in the integration and use of (multi-)omics data for a better understanding of the molecular mechanisms, processes, and pathways characterizing complex diseases like cancer. The success of such a holistic model in translational research and patient care depends on the gradual shift to a comprehensive systems approach, sustained by data sharing across and between different fields of expertise, accompanied by corresponding transformations in the political and regulatory environment to foster these developments.

Taken together, these endeavors have the potential to lead to innovative measures for disease prevention, early diagnosis, disease monitoring, and real-time decision-making, thus making precision medicine a forthcoming reality.

List of Figures

List of Publications

Peer-reviewed journal publications

Schneider, L., Kehl, T., Thedinga, K., Grammes, N. L., Backes, C., Mohr, C., Schubert, B., Lenhof, K., Gerstner, N., Hartkopf, A. D., Wallwiener, M., Kohlbacher, O., Keller, A., Meese, E., Graf, N., and Lenhof, H.-P. **ClinOmicsTrail^{bc}: a visual analytics tool for breast cancer treatment stratification using multi-omics data**. *Bioinformatics* (2019) 35.24. doi: 10.1093/bioinformatics/btz302

Schneider, L., Stöckel, D., Kehl, T., Gerasch, A., Ludwig, N., Leidinger, P., Huwer, H., Tenzer, S., Kohlbacher, O., Hildebrandt, A., Kaufmann, M., Gessler, M., Keller, A., Meese, E., Graf, N., and Lenhof, H.-P. **DrugTargetInspector: an assistance tool for patient treatment stratification**. *International Journal of Cancer* (2016) 138.7. doi: 10.1002/ijc.29897

Kehl, T., **Schneider, L.**, Kattler, K., Stöckel, D., Wegert, J., Gerstner, N., Ludwig, N., Distler, U., Tenzer, S., Gessler, M., Walter, J., Keller, A., Graf, N., Meese, E., and Lenhof, H.-P. **The role of TCF3 as potential master regulator in blastemal Wilms tumors**. *International Journal of Cancer* (2019) 144.6. doi: 10.1002/ijc.31834

Kehl, T., **Schneider, L.**, Kattler, K., Stöckel, D., Wegert, J., Gerstner, N., Ludwig, N., Distler, U., Schick, M., Keller, U., Tenzer, S., Gessler, M., Walter, J., Keller, A., Graf, N., Meese, E., and Lenhof, H.-P. **REGGAE: a novel approach for the identification of key transcriptional regulators**. *Bioinformatics* (2018) 1.8. doi: 10.1093/bioinformatics/bty372

Kehl, T., **Schneider, L.**, Schmidt, F., Stöckel, D., Gerstner, N., Backes, C., Meese, E., Keller, A., Schulz, M. H., and Lenhof, H.-P. **RegulatorTrail: a web service for the identification of key transcriptional regulators**. *Nucleic Acids Research* (2017) 45.W1. doi: 10.1093/nar/gkx350

Stöckel, D., Kehl, T., Trampert, P., **Schneider, L.**, Backes, C., Ludwig, N., Gerasch, A., Kaufmann, M., Gessler, M., Graf, N., Meese, E., Keller, A., and Lenhof, H.-P. **Multi-omics enrichment analysis using the GeneTrail2 web service**. *Bioinformatics* (2016) 31.10. doi: 10.1093/bioinformatics/btv770

*Biederstädt, A., *Hassan, Z., *Schneeweis, C., *Schick, M., **Schneider, L.**, Muckenhuber, A., Hong, Y., Nilsson, L., Wirth, M., Dantes, Z., Steiger, K., Schunck, K., Langston, S., Lenhof, H.-P., Coluccio, A., Orben, F., Slawska, J., Scherger, A.K., Saur, D., Müller, S., Rad, R., Weichert, W., Nilsson, J., Reichert, M., Schneider, G., and Keller, U. **SUMO Pathway Inhibition Targets an Aggressive Pancreatic Cancer Subtype**. *Gut* (2020) doi: 10.1136/gutjnl-2018-317856
* equally contributing first authors

Backes, C., Kehl, T., Stöckel, D., Fehlmann, T., **Schneider, L.**, Meese, E., Lenhof, H.-P., and Keller, A. **miRPathDB: a new dictionary on microRNAs and target pathways**. *Nucleic Acids Research* (2017) 45.D1. doi: 10.1093/nar/gkw926

Conference posters

Schneider, L., Stöckel, D., Kehl, T., Gerasch, A., Kaufmann, M., Kohlbacher, O., Keller, A., and Lenhof, H.-P. **DrugTargetInspector: an assistance tool for patient treatment stratification**. *ECCB 2016 The Hague* (*the 15th European Conference on Computational Biology*), The Hague, The Netherlands, September 3-September 7, 2016. doi: 10.7490/f1000research.1112901.1

Schneider, L., Stöckel, D., Kehl, T., Gerasch, A., Kaufmann, M., Kohlbacher, O., Keller, A., and Lenhof, H.-P. **DrugTargetInspector: an assistance tool for patient treatment stratification**. *Personalized Medicine Congress*, Tübingen, Germany, May 18-May 20, 2016.

Stöckel, D., Kehl, T., **Schneider, L.**, Müller, O., Backes, C., Rurainski, A., Gerasch, A., Keller, A., Kaufmann, M., and Lenhof, H.-P. **NetworkTrail: Network Analysis Toolbox**. *ISMB ECCB 2013 Berlin* (*the 21st Annual International Conference on Intelligent Systems for Molecular Biology and 12th European Conference on Computational Biology*), Berlin, Germany, July 21-July 23, 2013.

Publications in preparation

Schick, M., Maurer, S., **Schneider, L.**, Schunck, K., Rohleder, E., Maurer, C., Hofstetter, J., Baluapuri, A., Scherger, A. K., Slotta-Huspenina, J., Weber, J., Engleitner, T., Maresch, R., Slawska, J., Lewis, R., Istvanffy, R., Habringer, S., Steiger, K., Baiker, A., Oostendorp, R., Miething, C., Lenhof, H.-P., Chapuy, B., Wolf, E., Rad, R., Müller, S., and Keller, U. **The SUMO isopeptidase SENP6 controls chromatin dynamics to maintain genome integrity and suppress lymphomagenesis**. (manuscript under submission)

References

[1] Weinberg, R. A. *The biology of cancer*. Garland Science, second edition, 2014.

[2] Smith, B. D., Smith, G. L., Hurria, A., et al. Future of Cancer Incidence in the United States: Burdens Upon an Aging, Changing Nation. *Journal of Clinical Oncology*, 27(17):2758–2765, September 2016.

[3] Pedersen, J. K., Engholm, G., Skytthe, A., et al. Cancer and aging: Epidemiology and methodological challenges. *Acta Oncologica*, 55(sup1):7–12, January 2016.

[4] Siegel, R. L., Miller, K. D., and Jemal, A. Cancer statistics, 2016. *CA: A Cancer Journal for Clinicians*, 66(1):7–30, January 2016.

[5] Statistisches Bundesamt. Todesursachen in Deutschland 2015. Technical report, 2017.

[6] Burrell, R. A., McGranahan, N., Bartek, J., et al. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338–345, September 2013.

[7] Jouanna, J. The legacy of the Hippocratic treatise of the nature of man: The theory of the four humours. In *Greek Medicine from Hippocrates to Galen*, pages 335–360. jstor.org, 2012.

[8] Lièvre, A., Bachet, J.-B., Le Corre, D., et al. KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. *Cancer Research*, 66(8):3992–3995, April 2006.

[9] Sosman, J. A., Kim, K. B., Schuchter, L., et al. Survival in BRAF V600–Mutant Advanced Melanoma Treated with Vemurafenib. *New England Journal of Medicine*, 366(8):707–714, February 2012.

[10] Hauschild, A., Grob, J. J., Demidov, L. V., et al. Dabrafenib in BRAF-mutated metastatic melanoma: a multicentre, open-label, phase 3 randomised controlled trial. *The Lancet*, 380(9839):358–365, July 2012.

[11] Hyman, D. M., Puzanov, I., Subbiah, V., et al. Vemurafenib in Multiple Nonmelanoma Cancers with BRAF V600 Mutations. *New England Journal of Medicine*, 373(8):726–736, August 2015.

[12] Iyer, G., Hanrahan, A. J., Milowsky, M. I., et al. Genome Sequencing Identifies a Basis for Everolimus Sensitivity. *Science*, 338(6104):221–221, October 2012.

[13] Prasad, V. Perspective: The precision-oncology illusion. *Nature*, 537(7619):S63–S63, September 2016.

[14] Alyass, A., Turcotte, M., and Meyre, D. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Medical Genomics*, 8(1):613, 2015.

[15] Free Icons from Icons8. Available at https://icons8.com.

[16] Mendel, G. *Experiments in Plant Hybridisation*. 1866.

[17] Berg, J. M., Tymoczko, J. L., and Stryer, L. DNA, RNA, and the Flow of Genetic Information. In *Biochemistry*. W H Freeman, 2002.

[18] Bianconi, E., Piovesan, A., Facchin, F., et al. An estimation of the number of cells in the human body. *Annals of Human Biology*, 40(6):463–471, November 2013.

[19] Aben, N., Vis, D. J., Michaut, M., et al. TANDEM: a two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. *Bioinformatics*, 32(17):i413–i420, September 2016.

[20] Peterson, C. L. and Laniel, M.-A. Histones and histone modifications. *Current Biology*, 14(14):R546–R551, July 2004.

[21] Mitchell, P. J. and Tjian, R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*, 245(4916):371–378, July 1989.

[22] von Hippel, P. H. An Integrated Model of the Transcription Complex in Elongation, Termination, and Editing. *Science*, 281(5377):660–665, July 1998.

[23] Lesch, B. J. and Page, D. C. Poised chromatin in the mammalian germ line. *Development*, 141(19):3619–3626, September 2014.

[24] Struhl, K. Histone acetylation and transcriptional regulatory mechanisms. *Genes & Development*, 12(5):599–606, March 1998.

[25] Curradi, M., Izzo, A., Badaracco, G., et al. Molecular Mechanisms of Gene Silencing Mediated by DNA Methylation. *Molecular and Cellular Biology*, 22(9):3157–3173, May 2002.

[26] Nilsen, T. W. and Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–463, January 2010.

[27] Proudfoot, N. J., Furger, A., and Dye, M. J. Integrating mRNA Processing with Transcription. *Cell*, 108(4):501–512, February 2002.

[28] Gott, J. M. and Emeson, R. B. Functions and Mechanisms of RNA Editing. *Annual Review of Genetics*, 34(1):499–531, December 2000.

[29] Mattick, J. S. and Makunin, I. V. Non-coding RNA. *Human Molecular Genetics*, 15:R17–R29, April 2006.

[30] Cai, Y., Yu, X., Hu, S., et al. A Brief Review on the Mechanisms of miRNA Regulation. *Genomics, Proteomics & Bioinformatics*, 7(4):147–154, December 2009.

[31] Green, R. and Noller, H. F. Ribosomes and Translation. *Annual Review of Biochemistry*, 66(1):679–716, June 1997.

[32] Hartl, F. U. and Hayer-Hartl, M. Molecular Chaperones in the Cytosol: from Nascent Chain to Folded Protein. *Science*, 295(5561):1852–1858, March 2002.

[33] Walsh, C. *Posttranslational Modification of Proteins*. Expanding Nature's Inventory. W. H. Freeman, 2006.

[34] Pickart, C. M. Mechanisms Underlying Ubiquitination. *Annual Review of Biochemistry*, 70(1):503–533, June 2001.

[35] Flotho, A. and Melchior, F. Sumoylation: A Regulatory Protein Modification in Health and Disease. *Annual Review of Biochemistry*, 82(1):357–385, June 2013.

[36] Schlessinger, J. Cell Signaling by Receptor Tyrosine Kinases. *Cell*, 103(2):211–225, October 2000.

[37] Kitano, H. Biological robustness. *Nature Reviews Genetics*, 5(11):826–837, November 2004.

[38] Oeppen, J. and Vaupel, J. W. Broken Limits to Life Expectancy. *Science*, 296(5570):1029–1031, May 2002.

[39] Cao, B., Bray, F., Beltrán-Sánchez, H., et al. Benchmarking life expectancy and cancer mortality: global comparison with cardiovascular disease 1981-2010. *BMJ*, 357:j2765, June 2017.

[40] Loeb, L. A. Human Cancers Express a Mutator Phenotype: Hypothesis, Origin, and Consequences. *Cancer Research*, 76(8):2057–2059, April 2016.

[41] Turnell, A. S. and Grand, R. J. DNA viruses and the cellular DNA-damage response. *Journal of General Virology*, 93(10):2076–2097, October 2012.

[42] Stratton, M. R., Campbell, P. J., and Futreal, P. A. The cancer genome. *Nature*, 458(7239):719–724, April 2009.

[43] Rahman, N. and Stratton, M. R. The Genetics of Breast Cancer Susceptibility. *Annual Review of Genetics*, 32(1):95–121, December 1998.

[44] de la Chapelle, A. Genetic predisposition to colorectal cancer. *Nature Reviews Cancer*, 4(10):769–780, October 2004.

[45] Garber, J. E. and Offit, K. Hereditary Cancer Predisposition Syndromes. *Journal of Clinical Oncology*, 23(2):276–292, September 2016.

[46] Vogelstein, B. and Kinzler, K. W. The multistep nature of cancer. *Trends in Genetics*, 9(4):138–141, April 1993.

[47] Solomon, E., Borrow, J., and Goddard, A. Chromosome aberrations and cancer. *Science*, 254(5035):1153–1160, November 1991.

[48] Ducasse, M. Epigenetic aberrations and cancer. *Molecular Cancer*, 5(1):60, 2006.

[49] Hanahan, D. and Weinberg, R. A. The Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, March 2011.

[50] Fedi, P., Tronick, S. R., and Aaronson, S. A. Growth factors. In *Cancer Medicine*, pages 41–64. Williams and Wilkins, 1999.

[51] Weinberg, R. A. The retinoblastoma protein and cell cycle control. *Cell*, 81(3):323–330, May 1995.

[52] Jerry W Shay, W. E. W. Role of telomeres and telomerase in cancer. *Seminars in Cancer Biology*, 21(6):349–353, December 2011.

[53] Pavlova, N. N. and Thompson, C. B. The Emerging Hallmarks of Cancer Metabolism. *Cell Metabolism*, 23(1):27–47, January 2016.

[54] Wright, W. E., Pereira-Smith, O. M., and Shay, J. W. Reversible cellular senescence: implications for immortalization of normal human diploid fibroblasts. *Molecular and Cellular Biology*, 9(7):3088–3092, July 1989.

[55] Fernald, K. Evading apoptosis in cancer. *Trends in Cell Biology*, 23(12):620–633, December 2013.

[56] Vinay, D. S., Ryan, E. P., Pawelec, G., et al. Immune evasion in cancer: Mechanistic basis and therapeutic strategies. *Seminars in Cancer Biology*, 35:S185–S198, December 2015.

[57] Patel, S. P. and Kurzrock, R. PD-L1 Expression as a Predictive Biomarker in Cancer Immunotherapy. *Molecular Cancer Therapeutics*, 14(4):molcanther.0983.2014–856, February 2015.

[58] Bouck, N., Stellmach, V., and Hsu, S. C. How Tumors Become Angiogenic. *Advances in Cancer Research*, 69:135–174, January 1996.

[59] Sporn, M. B. The war on cancer. *The Lancet*, 347(9012):1377–1381, May 1996.

[60] Lengauer, C., Kinzler, K. W., and Vogelstein, B. Genetic instabilities in human cancers. *Nature*, 396(6712):643–649, December 1998.

[61] Grivennikov, S. I., Greten, F. R., and Karin, M. Immunity, Inflammation, and Cancer. *Cell*, 140(6):883–899, March 2010.

[62] Coffey, J. C., Wang, J. H., Smith, M., et al. Excisional surgery for cancer cure: therapy at a cost. *The Lancet. Oncology*, 4(12):760–768, December 2003.

[63] Baskar, R., Dai, J., Wenlong, N., et al. Biological response of cancer cells to radiation treatment. *Frontiers in Molecular Biosciences*, 1:834, 2014.

[64] American Cancer Society. A to Z List of Cancer Drugs. www.cancer.gov/about-cancer/treatment/drugs.

[65] Corrie, P. G. Cytotoxic chemotherapy: clinical aspects. *Medicine*, 39(12):717–722, December 2011.

[66] Weinstein, I. B., Joe, A., and Felsher, D. Oncogene Addiction. *Cancer Research*, 68(9):3077–3080, May 2008.

[67] Esteva, F. J. Monoclonal Antibodies, Small Molecules, and Vaccines in the Treatment of Breast Cancer. *The Oncologist*, 9(Supplement 3):4–9, June 2004.

[68] Wu, P., Nielsen, T. E., and Clausen, M. H. Small-molecule kinase inhibitors: an analysis of FDA-approved drugs. *Drug Discovery Today*, 21(1):5–10, January 2016.

[69] Jabbour, E., Fava, C., and Kantarjian, H. Advances in the biology and therapy of patients with chronic myeloid leukaemia. *Best Practice & Research. Clinical Haematology*, 22(3):395–407, September 2009.

[70] Chapman, P. B., Hauschild, A., Robert, C., et al. Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation. *New England Journal of Medicine*, 364(26):2507–2516, June 2011.

[71] Vogel, C. L., Cobleigh, M. A., Tripathy, D., et al. Efficacy and Safety of Trastuzumab as a Single Agent in First-Line Treatment of HER2-Overexpressing Metastatic Breast Cancer. *Journal of Clinical Oncology*, 20(3):719–726, February 2002.

[72] Ranieri, G., Patruno, R., Ruggieri, E., et al. Vascular Endothelial Growth Factor (VEGF) as a Target of Bevacizumab in Cancer: From the Biology to the Clinic. *Current Medicinal Chemistry*, 13(16):1845–1857, June 2006.

[73] Coiffier, B., Lepage, E., Brière, J., et al. CHOP Chemotherapy plus Rituximab Compared with CHOP Alone in Elderly Patients with Diffuse Large-B-Cell Lymphoma. *New England Journal of Medicine*, 346(4):235–242, January 2002.

[74] Diamantis, N. and Banerji, U. Antibody-drug conjugates—an emerging class of cancer treatment. *British Journal of Cancer*, 114(4):362–367, January 2016.

[75] Strebhardt, K. and Ullrich, A. Paul Ehrlich's magic bullet concept: 100 years of progress. *Nature Reviews Cancer*, 8(6):473–480, June 2008.

[76] Huang, M., Shen, A., Ding, J., et al. Molecularly targeted cancer therapy: some lessons from the past decade. *Trends in Pharmacological Sciences*, 35(1):41–50, January 2014.

[77] Longley, D. B. and Johnston, P. G. Molecular mechanisms of drug resistance. *The Journal of Pathology*, 205(2):275–292, January 2005.

[78] Rimawi, M. F. and Osborne, C. K. Breast Cancer: Blocking both driver and escape pathways improves outcomes. *Nature Reviews Clinical Oncology*, 9(3):133–134, March 2012.

[79] Yap, T. A., Carden, C. P., and Kaye, S. B. Beyond chemotherapy: targeted therapies in ovarian cancer. *Nature Reviews Cancer*, 9(3):167–181, March 2009.

[80] Luo, J., Solimini, N. L., and Elledge, S. J. Principles of cancer therapy: oncogene and non-oncogene addiction. *Cell*, 136(5):823–837, March 2009.

[81] Kummar, S., Chen, H. X., Wright, J., et al. Utilizing targeted cancer therapeutic agents in combination: novel approaches and urgent requirements. *Nature Reviews Drug Discovery*, 9(11):843–856, November 2010.

[82] Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nature Reviews Cancer*, 12(4):252–264, April 2012.

[83] Le, D. T., Uram, J. N., Wang, H., et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *New England Journal of Medicine*, 372(26):2509–2520, June 2015.

[84] Inman, B. A., Longo, T. A., Ramalingam, S., et al. Atezolizumab: A PD-L1–Blocking Antibody for Bladder Cancer. *Clinical Cancer Research*, 23(8):1886–1890, April 2017.

[85] Yarchoan, M., Hopkins, A., and Jaffee, E. M. Tumor Mutational Burden and Response Rate to PD-1 Inhibition. *New England Journal of Medicine*, 377(25):2500–2501, December 2017.

[86] Goodman, A. M., Kato, S., Bazhenova, L., et al. Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers. *Molecular Cancer Therapeutics*, 16(11):2598–2608, November 2017.

[87] June, C. H. Principles of adoptive T cell cancer therapy. *The Journal of Clinical Investigation*, 117(5):1204–1212, May 2007.

[88] Ott, P. A., Hu, Z., Keskin, D. B., et al. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature*, 547(7662):217–221, July 2017.

[89] Sahin, U., Derhovanessian, E., Miller, M., et al. Personalized RNA mutanome vaccines mobilize polyspecific therapeutic immunity against cancer. *Nature*, 547(7662):222–226, July 2017.

[90] Fu, J., Malm, I.-J., Kadayakkara, D. K., et al. Preclinical evidence that PD-1 blockade cooperates with cancer vaccine TEGVAX to elicit regression of established tumors. *Cancer Research*, 74(15), May 2014.

[91] Nowell, P. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, October 1976.

[92] Naghavi, M., Abajobir, A. A., Abbafati, C., et al. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, 390(10100):1151–1210, 2017.

[93] Redekop, W. K. and Mladsi, D. The Faces of Personalized Medicine: A Framework for Understanding Its Meaning and Scope. *Value in Health*, 16(6):S4–S9, September 2013.

[94] Metzker, M. L. Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1):31–46, December 2009.

[95] Bumgarner, R. DNA microarrays: Types, Applications and their future. *Current Protocols in Molecular Biology*, January 2013.

[96] Bantscheff, M., Schirle, M., Sweetman, G., et al. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and Bioanalytical Chemistry*, 389(4):1017–1031, August 2007.

[97] Jain, K. K. *Textbook of Personalized Medicine*. Springer, second edition, 2015.

[98] Hood, L. and Flores, M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *New Biotechnology*, 29(6):613–624, September 2012.

[99] King, M.-C., Marks, J. H., Mandell, J. B., et al. Breast and Ovarian Cancer Risks Due to Inherited Mutations in BRCA1 and BRCA2. *Science*, 302(5645):643–646, October 2003.

[100] Warner, E. Surveillance of BRCA1 and BRCA2 Mutation Carriers With Magnetic Resonance Imaging, Ultrasound, Mammography, and Clinical Breast Examination. *JAMA*, 292(11):1317–1325, September 2004.

[101] Hartmann, L. C., Sellers, T. A., Schaid, D. J., et al. Efficacy of Bilateral Prophylactic Mastectomy in BRCA1 and BRCA2 Gene Mutation Carriers. *Journal of the National Cancer Institute*, 93(21):1633–1637, November 2001.

[102] Pruthi, S., Heisey, R. E., and Bevers, T. B. Chemoprevention for Breast Cancer. *Annals of Surgical Oncology*, 22(10):3230–3235, 2015.

[103] F C Gaertner, M. S. G. B. and Beer, A. J. Imaging of Hypoxia Using PET and MRI. *Current Pharmaceutical Biotechnology*, 13(4):552–570, February 2012.

[104] Leslie J Sheffield, H. E. P. Clinical Use of Pharmacogenomic Tests in 2009. *The Clinical Biochemist Reviews*, 30(2):55, May 2009.

[105] FDA-NIH Biomarker Working Group. *BEST (Biomarkers, EndpointS, and other Tools)*. Food and Drug Administration (US), 2016.

[106] Paik, S., Shak, S., Tang, G., et al. A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. *New England Journal of Medicine*, 351(27):2817–2826, December 2004.

[107] Slodkowska, E. A. and Ross, J. S. MammaPrint™ 70-gene signature: another milestone in personalized medical care for breast cancer patients. *Expert Review of Molecular Diagnostics*, 9(5):417–422, January 2014.

[108] Parkinson, D. R., Johnson, B. E., and Sledge, G. W. Making Personalized Cancer Medicine a Reality: Challenges and Opportunities in the Development of Biomarkers and Companion Diagnostics. *Clinical Cancer Research*, 18(3):619–624, February 2012.

[109] Vogel, C. L., Cobleigh, M. A., Tripathy, D., et al. Efficacy and Safety of Trastuzumab as a Single Agent in First-Line Treatment of HER2-Overexpressing Metastatic Breast Cancer. *Journal of Clinical Oncology*, 20(3):719–726, September 2016.

[110] Loree, J. M., Kopetz, S., and Raghav, K. P. S. Current companion diagnostics in advanced colorectal cancer; getting a bigger and better piece of the pie. *Journal of Gastrointestinal Oncology*, 8(1):199–212, February 2017.

[111] Kantarjian, H. M., Cortes, J., O'Brien, S., et al. Imatinib mesylate (STI571) therapy for Philadelphia chromosome–positive chronic myelogenous leukemia in blast phase. *Blood*, 99(10):3547–3553, May 2002.

[112] Kantarjian, H., Giles, F., Wunderle, L., et al. Nilotinib in Imatinib-Resistant CML and Philadelphia Chromosome–Positive ALL. *New England Journal of Medicine*, 354(24):2542–2551, June 2006.

[113] Ashburn, T. T. and Thor, K. B. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3(8):673–683, August 2004.

[114] Sleigh, S. H. and Barton, C. L. Repurposing Strategies for Therapeutics. *Pharmaceutical Medicine*, 24(3):151–159, August 2012.

[115] Paez, J. G., Jänne, P. A., Lee, J. C., et al. EGFR Mutations in Lung Cancer: Correlation with Clinical Response to Gefitinib Therapy. *Science*, 304(5676):1497–1500, June 2004.

[116] Camidge, D. R., Bang, Y.-J., Kwak, E. L., et al. Activity and safety of crizotinib in patients with ALK-positive non-small-cell lung cancer: updated results from a phase 1 study. *The Lancet. Oncology*, 13(10):1011–1019, October 2012.

[117] Mosse, Y. P., Balis, F. M., Lim, M. S., et al. Efficacy of crizotinib in children with relapsed/refractory ALK-driven tumors including anaplastic large cell lymphoma and neuroblastoma: A Children's Oncology Group phase I consortium study. In *2012 ASCO Annual Meeting*. 2012.

[118] Lemery, S., Keegan, P., and Pazdur, R. First FDA Approval Agnostic of Cancer Site — When a Biomarker Defines the Indication. *New England Journal of Medicine*, 377(15):1409–1412, October 2017.

[119] Massard, C., Michiels, S., Ferté, C., et al. High-Throughput Genomics and Clinical Outcome in Hard-to-Treat Advanced Cancers: Results of the MOSCATO 01 Trial. *Cancer Discovery*, 7(6):586–595, June 2017.

[120] Conley, B. A., Gray, R., Chen, A., et al. Abstract CT101: NCI-molecular analysis for therapy choice (NCI-MATCH) clinical trial: interim analysis:. *Cancer Research*, 76(14 Supplement):CT101–CT101, July 2016.

[121] West, M., Ginsburg, G. S., Huang, A. T., et al. Embracing the complexity of genomic data for personalized medicine. *Genome Research*, 16(5):559–566, May 2006.

[122] Hamburg, M. A. and Collins, F. S. The Path to Personalized Medicine. *New England Journal of Medicine*, 363(4):301–304, July 2010.

[123] Yadav, S. P. The Wholeness in Suffix -omics, -omes, and the Word Om. *Journal of Biomolecular Techniques*, 18(5):277, December 2007.

[124] 3D-Gene®. Outline of detection method of genes by DNA microarrays. http://www.3d-gene.com/en/about/chip/chi_003.html .

[125] Sanborn, M. E., Connolly, B. K., Gurunathan, K., et al. Fluorescence Properties and Photophysics of the Sulfoindocyanine Cy3 Linked Covalently to DNA. *The Journal of Physical Chemistry B*, 111(37):11064–11074, September 2007.

[126] Zhang, D. Y., Chen, S. X., and Yin, P. Optimizing the specificity of nucleic acid hybridization. *Nature Chemistry*, 4(3):208–214, January 2012.

[127] Shalon, D., Smith, S. J., and Brown, P. O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research*, 6(7):639–645, July 1996.

[128] Kallioniemi, A., Kallioniemi, O. P., Sudar, D., et al. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, 258(5083):818–821, October 1992.

[129] Van Prooijen-Knegt, A. C., Van Hoek, J. F. M., Bauman, J. G. J., et al. In situ hybridization of DNA sequences in human metaphase chromosomes visualized by an indirect fluorescent immunocytochemical procedure. *Experimental Cell Research*, 141(2):397–407, October 1982.

[130] Kallioniemi, O.-P., Kallioniemi, A., Piper, J., et al. Optimizing comparative genomic hybridization for analysis of DNA sequence copy number changes in solid tumors. *Genes, Chromosomes and Cancer*, 10(4):231–243, August 1994.

[131] Schumacher, A., Kapranov, P., Kaminsky, Z., et al. Microarray-based DNA methylation profiling: technology and applications. *Nucleic Acids Research*, 34(2):528–542, January 2006.

[132] Lister, R. and Ecker, J. R. Finding the fifth base: Genome-wide sequencing of cytosine methylation. *Genome Research*, 19(6):959–966, June 2009.

[133] Frommer, M., McDonald, L. E., Millar, D. S., et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America*, 89(5):1827–1831, March 1992.

[134] Bibikova, M., Barnes, B., Tsan, C., et al. High density DNA methylation array with single CpG site resolution. *Science*, 98(4):288–295, October 2011.

[135] Hartmann, M., Roeraade, J., Stoll, D., et al. Protein microarrays for diagnostic assays. *Analytical and Bioanalytical Chemistry*, 393(5):1407–1416, September 2008.

[136] Holmes, K. L. and Lantz, L. M. Protein labeling with fluorescent probes. *Methods in Cell Biology*, 63:185–204, January 2001.

[137] Zong, Y., Zhang, S., Chen, H.-T., et al. Forward-Phase and Reverse-Phase Protein Microarray . *Microarrays*, pages 363–373, 2007.

[138] Spurrier, B., Ramalingam, S., and Nishizuka, S. Reverse-phase protein lysate microarrays for cell signaling analysis. *Nature Protocols*, 3(11):1796–1808, November 2008.

[139] Brody, E. N., Gold, L., Lawn, R. M., et al. High-content affinity-based proteomics: unlocking protein biomarker discovery. *Expert Review of Molecular Diagnostics*, 10(8):1013–1022, January 2014.

[140] SomaLogic, Inc. SOMAscan® Proteomic Assay Technical White Paper . Technical report, 2016.

[141] Huang, Y. and Zhu, H. Protein Array-based Approaches for Biomarker Discovery in Cancer. *Genomics, Proteomics & Bioinformatics*, 15(2):73–81, April 2017.

[142] Bajcsy, P. An Overview of DNA Microarray Grid Alignment and Foreground Separation Approaches. *EURASIP Journal on Advances in Signal Processing*, 2006(1), April 2006.

[143] Adams, R. and Bischof, L. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6):641–647, June 1994.

[144] Bozinov, D. and Rahnenführer, J. Unsupervised technique for robust target separation and analysis of DNA microarray spots through adaptive pixel clustering. *Bioinformatics*, 18(5), May 2002.

[145] Reilly, C., Raghavan, A., and Bohjanen, P. Global Assessment of Cross-Hybridization for Oligonucleotide Arrays. *Journal of Biomolecular Techniques*, 17(2):163, April 2006.

[146] Hoaglin, D. C., Mosteller, F., and Tukey, J. W. *Understanding Robust and Exploratory Data Analysis*. Wiley, June 2000.

[147] Opitz, L., Salinas-Riester, G., Grade, M., et al. Impact of RNA degradation on gene expression profiling. *BMC Medical Genomics*, 3(1):530, August 2010.

[148] Steger, D., Berry, D., Haider, S., et al. Systematic Spatial Bias in DNA Microarray Hybridization Is Caused by Probe Spot Position-Dependent Variability in Lateral Diffusion. *PLOS ONE*, 6(8), August 2011.

[149] Stafford, P. *Methods in Microarray Normalization*. CRC Press, 2008.

[150] Rydén, P., Andersson, H., Landfors, M., et al. Evaluation of microarray data normalization procedures using spike-in experiments. *BMC Bioinformatics*, 7(1):300, 2006.

[151] Bolstad, B. M., Irizarry, R. A., Astrand, M., et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, January 2003.

[152] Durbin, B. P., Hardin, J. S., Hawkins, D. M., et al. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18:S105–S110, July 2002.

[153] Irizarry, R. A., Hobbs, B., Collin, F., et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, April 2003.

[154] Hubbell, E., Liu, W. M., and Mei, R. Robust estimators for expression analysis. *Bioinformatics*, 18(12):1585–1592, December 2002.

[155] Leek, J. T., Scharpf, R. B., Bravo, H. C., et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, October 2010.

[156] Brown, J. S., Kuhn, D., Wisser, R., et al. Quantification of sources of variation and accuracy of sequence discrimination in a replicated microarray experiment. *BioTechniques*, 36(2):324–332, February 2004.

[157] Lazar, C., Meganck, S., Taminau, J., et al. Batch effect removal methods for microarray gene expression data integration: a survey. *Briefings in Bioinformatics*, 14(4):469–490, July 2013.

[158] Leek, J. T. and Storey, J. D. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLOS Genetics*, 3(9), 2007.

[159] Johnson, W. E., Li, C., and Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, January 2007.

[160] Gagnon-Bartsch, J. A. and Speed, T. P. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, July 2012.

[161] Chen, C., Grennan, K., Badner, J., et al. Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. *PLOS ONE*, 6(2), February 2011.

[162] Sanger, F. and Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3):441–448, May 1975.

[163] Shendure, J. and Ji, H. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145, October 2008.

[164] Schadt, E. E., Turner, S., and Kasarskis, A. A window into third-generation sequencing. *Human Molecular Genetics*, 19(R2):R227–R240, October 2010.

[165] Ambardar, S., Gupta, R., Trakroo, D., et al. High Throughput Sequencing: An Overview of Sequencing Chemistry. *Indian Journal of Microbiology*, 56(4):394–404, July 2016.

[166] Fuller, C. W., Middendorf, L. R., Benner, S. A., et al. The challenges of sequencing by synthesis. *Nature Biotechnology*, 27(11):1013–1023, November 2009.

[167] Mirzabekov, A. D. DNA sequencing by hybridization — a megasequencing method and a diagnostic tool? *Trends in Biotechnology*, 12(1):27–32, January 1994.

[168] Heather, J. M. and Chain, B. The sequence of sequencers: The history of sequencing DNA. *Science*, 107(1):1–8, January 2016.

[169] Rothberg, J. M., Hinz, W., Rearick, T. M., et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–352, July 2011.

[170] Jain, M., Olsen, H. E., Paten, B., et al. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome biology*, 17(1):239, December 2016.

[171] Illumina. Illumina Sequencing Technology. https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf. Technical report, 2010.

[172] Snipcademy. Illumina Sequencing-By-Synthesis (SBS) Technology. https://binf.snipcademy.com/lessons/ngs-techniques/illumina-solexa, 2017.

[173] Knierim, E., Lucke, B., Schwarz, J. M., et al. Systematic Comparison of Three Methods for Fragmentation of Long-Range PCR Products for Next Generation Sequencing. *PLOS ONE*, 6(11), November 2011.

[174] Bronner, I. F., Quail, M. A., Turner, D. J., et al. Improved Protocols for Illumina Sequencing. *Current Protocols in Human Genetics*, 0(18), July 2009.

[175] Illumina. Data Sheet: Genomic Sequencing. https://www.illumina.com/Documents/products/datasheets/datasheet_genomic_sequence.pdf. Technical report, 2010.

[176] Quinlan, A. R., Boland, M. J., Leibowitz, M. L., et al. Genome Sequencing of Mouse Induced Pluripotent Stem Cells Reveals Retroelement Stability and Infrequent DNA Rearrangement during Reprogramming. *Cell Stem Cell*, 9(4):366–373, October 2011.

[177] Koboldt, D. C., Chen, K., Wylie, T., et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17):2283–2285, September 2009.

[178] Ng, S. B., Turner, E. H., Robertson, P. D., et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272–276, August 2009.

[179] Iglesias, A., Anyane-Yeboa, K., Wynn, J., et al. The usefulness of whole-exome sequencing in routine clinical practice. *Genetics in Medicine*, 16(12):922–931, December 2014.

[180] Warr, A., Robert, C., Hume, D., et al. Exome Sequencing: Current and Future Perspectives. *G3: Genes, Genomes, Genetics*, 5(8):1543–1550, August 2015.

[181] Rehm, H. L. Disease-targeted sequencing: a cornerstone in the clinic. *Nature Reviews Genetics*, 14(4):295–300, April 2013.

[182] Shin, H.-T., Choi, Y.-L., Yun, J. W., et al. Prevalence and detection of low-allele-fraction variants in clinical cancer samples. *Nature Communications*, 8(1):1377, November 2017.

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

[184] 't Hoen, P. A. C., Ariyurek, Y., Thygesen, H. H., et al. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Research*, 36(21):e141–e141, December 2008.

[185] Laird, P. W. Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics*, 11(3):191–203, March 2010.

[186] Soozangar, N., Sadeghi, M. R., Jeddi, F., et al. Comparison of genome-wide analysis techniques to DNA methylation analysis in human cancer. *Journal of Cellular Physiology*, 233(5):3968–3981, May 2018.

[187] Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, October 2009.

[188] Zambelli, F., Pesole, G., and Pavesi, G. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings in Bioinformatics*, 14(2):225–237, March 2013.

[189] Christian Ledergerber, C. D. Base-calling for next-generation sequencing platforms. *Briefings in Bioinformatics*, 12(5):489–497, September 2011.

[190] Sandmann, S., de Graaf, A. O., Karimi, M., et al. Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Scientific Reports*, 7(1):43169, February 2017.

[191] Pabinger, S., Dander, A., Fischer, M., et al. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*, 15(2):256–278, March 2014.

[192] Chandramohan, R., Wu, P.-Y., Phan, J. H., et al. Benchmarking RNA-Seq quantification tools. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 647–650. IEEE, 2013.

[193] ECSEQ Bioinformatics. Why does the per base sequence quality decrease over the read in Illumina? https://www.ecseq.com/support/ngs/why-does-the-sequence-quality-decrease-over-the-read-in-illumina, January 2017.

[194] Bolger, A. M., Lohse, M., and Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, August 2014.

[195] Dodt, M., Roehr, J., Ahmed, R., et al. FLEXBAR—Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology*, 1(3):895–905, December 2012.

[196] Venter, J. C., Adams, M. D., Myers, E. W., et al. The Sequence of the Human Genome. *Science*, 5(4):954–956, February 2001.

[197] Fonseca, N. A., Rung, J., Brazma, A., et al. Tools for mapping high-throughput sequencing data. *Bioinformatics*, 28(24):3169–3177, December 2012.

[198] Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009.

[199] Langmead, B. and Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, April 2012.

[200] Reinert, K., Langmead, B., Weese, D., et al. Alignment of Next-Generation Sequencing Reads. *Annual Review of Genomics and Human Genetics*, 16(1):133–151, August 2015.

[201] Kärkkäinen, J. and Sanders, P. Simple Linear Work Suffix Array Construction. In *Automata, Languages and Programming*. Springer, Berlin, Heidelberg, June 2003.

[202] Ferragina, P. and Manzini, G. *Opportunistic data structures with applications*. IEEE Computer Society, November 2000.

[203] Myers, E. W. An O(ND) difference algorithm and its variations. *Algorithmica*, 1(1-4):251–266, November 1986.

[204] Myers, G. A fast bit-vector algorithm for approximate string matching based on dynamic programming. In *Combinatorial Pattern Matching*, pages 1–13. Springer, Berlin, Heidelberg, Berlin, Heidelberg, July 1998.

[205] Kim, D., Pertea, G., Trapnell, C., et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):R36, April 2013.

[206] Petrov, D. A. Evolution of genome size: new approaches to an old problem. *Trends in Genetics*, 17(1):23–28, January 2001.

[207] McKenna, A., Hanna, M., Banks, E., et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, September 2010.

[208] Van der Auwera, G. A., Carneiro, M. O., Hartl, C., et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, 43(1):11.10.1–11.10.33, October 2013.

[209] Nielsen, R., Paul, J. S., Albrechtsen, A., et al. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, June 2011.

[210] Cibulskis, K., Lawrence, M. S., Carter, S. L., et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3):213–219, March 2013.

[211] Larson, D. E., Harris, C. C., Chen, K., et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28(3):311–317, December 2011.

[212] Koboldt, D. C., Zhang, Q., Larson, D. E., et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576, March 2012.

[213] Zhao, M., Wang, Q., Wang, Q., et al. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, 14(Suppl 11):S1, 2013.

[214] Guan, P. and Sung, W. K. Structural variation detection using next-generation sequencing data: A comparative technical review. *Methods*, 102:36–49, June 2016.

[215] Gerstung, M., Beisel, C., Rechsteiner, M., et al. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nature Communications*, 3(1):23, May 2012.

[216] Sun, J. X., He, Y., Sanford, E., et al. A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal. *PLoS computational biology*, 14(2), February 2018.

[217] Raphael, B. J., Dobson, J. R., Oesper, L., et al. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Medicine*, 6(1), 2014.

[218] Wang, K., Li, M., and Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):e164–e164, September 2010.

[219] Cingolani, P., Platts, A., Wang, L. L., et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, 6(2):80–92, October 2014.

[220] McLaren, W., Gil, L., Hunt, S. E., et al. The Ensembl Variant Effect Predictor. *Genome biology*, 17(1):122, December 2016.

[221] Kumar, P., Henikoff, S., and Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(7):1073–1081, July 2009.

[222] Adzhubei, I. A., Schmidt, S., Peshkin, L., et al. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249, April 2010.

[223] Landrum, M. J., Lee, J. M., Riley, G. R., et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(D1):D980–D985, January 2014.

[224] Griffith, M., Spies, N. C., Krysiak, K., et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature Genetics*, 49(2):170–174, February 2017.

[225] Tate, J. G., Bamford, S., Jubb, H. C., et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, 45:D777, October 2018.

[226] Sherry, S. T., Ward, M. H., Kholodov, M., et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, January 2001.

[227] Abyzov, A., Urban, A. E., Snyder, M., et al. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, 21(6):974–984, June 2011.

[228] Xie, C. and Tammi, M. T. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, 10(1):525, March 2009.

[229] Hormozdiari, F., Alkan, C., Eichler, E. E., et al. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Research*, 19(7):1270–1278, July 2009.

[230] Dobin, A., Davis, C. A., Schlesinger, F., et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, January 2013.

[231] Mortazavi, A., Williams, B. A., McCue, K., et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, July 2008.

[232] Anders, S., Pyl, P. T., and Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, January 2015.

[233] Liao, Y., Smyth, G. K., and Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, April 2014.

[234] FPKM. EMBL-EBI Glossary. www.ebi.ac.uk/training/online/glossary/fpkm.

[235] Conesa, A., Madrigal, P., Tarazona, S., et al. A survey of best practices for RNA-seq data analysis. *Genome biology*, 17(1):1, 2016.

[236] Trapnell, C., Roberts, A., Goff, L., et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7(3):562–578, March 2012.

[237] Anders, S. and Huber, W. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, October 2010.

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

[239] Elias, J. E. and Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4(3):207–214, March 2007.

[240] Niessen, W. M. A. *Liquid Chromatography-Mass Spectrometry*. CRC Press, August 2006.

[241] Fenn, J., Mann, M., Meng, C., et al. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, October 1989.

[242] Mitchell Wells, J. and McLuckey, S. A. Collision-Induced Dissociation (CID) of Peptides and Proteins. *Methods in Enzymology*, 402:148–185, January 2005.

[243] Cottrell, J. S. Protein identification using MS/MS data. *Journal of Proteomics*, 74(10):1842–1851, September 2011.

[244] Olsen, J. V. and Mann, M. Status of Large-scale Analysis of Post-translational Modifications by Mass Spectrometry. *Molecular & Cellular Proteomics*, 12(12):3444–3452, December 2013.

[245] Vogel, C. and Marcotte, E. M. Calculating absolute and relative protein abundance from mass spectrometry-based protein expression data. *Nature Protocols*, 3(9):1444–1451, September 2008.

[246] Taylor, P. J. Matrix effects: the Achilles heel of quantitative high-performance liquid chromatography–electrospray–tandem mass spectrometry. *Clinical Biochemistry*, 38(4):328–334, April 2005.

[247] Ong, S.-E., Blagoev, B., Kratchmarova, I., et al. Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Molecular & Cellular Proteomics*, 1(5):376–386, May 2002.

[248] Zhang, G. and Neubert, T. A. Use of Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC) for Phosphotyrosine Protein Identification and Quantitation. In *Phospho-Proteomics*, pages 79–92. Humana Press, Totowa, NJ, 2009.

[249] Wheatley, D. N., Scott, L., Lamb, J., et al. Single Amino Acid (Arginine) Restriction: Growth and Death of Cultured HeLa and Human Diploid Fibroblasts. *Cellular Physiology and Biochemistry*, 10(1-2):37–55, July 2000.

[250] Scott, L., Lamb, J., Smith, S., et al. Single amino acid (arginine) deprivation: rapid and selective death of cultured transformed and malignant cells. *British Journal of Cancer*, 83(6):800–810, September 2000.

[251] Geiger, T., Wisniewski, J. R., Cox, J., et al. Use of stable isotope labeling by amino acids in cell culture as a spike-in standard in quantitative proteomics. *Nature Protocols*, 6(2):147–157, February 2011.

[252] Palagi, P. M., Hernandez, P., Walther, D., et al. Proteome informatics I: Bioinformatics tools for processing experimental data. *Proteomics*, 6(20):5435–5444, October 2006.

[253] Allmer, J. Existing bioinformatics tools for the quantitation of post-translational modifications. *Amino Acids*, 42(1):129–138, May 2010.

[254] Mueller, L. N., Brusniak, M.-Y., Mani, D. R., et al. An Assessment of Software Solutions for the Analysis of Mass Spectrometry Based Quantitative Proteomics Data. *Journal of Proteome Research*, 7(1):51–61, January 2008.

[255] Röst, H. L., Sachsenberg, T., Aiche, S., et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nature Methods*, 13(9):741–748, September 2016.

[256] Pluskal, T., Castillo, S., Villar-Briones, A., et al. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, 11(1):395, 2010.

[257] Clasquin, M. F., Melamud, E., and Rabinowitz, J. D. LC-MS Data Processing with MAVEN: A Metabolomic Analysis and Visualization Engine. *Current Protocols in Bioinformatics*, 37(1):14.11.1–14.11.23, March 2012.

[258] Zhang, J.-F., He, S.-M., Cai, J.-J., et al. Preprocessing of Tandem Mass Spectrometric Data Based on Decision Tree Classification. *Genomics, Proteomics & Bioinformatics*, 3(4):231–237, 2005.

[259] Jaitly, N., Mayampurath, A., Littlefield, K., et al. Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data. *BMC Bioinformatics*, 10(1):87, 2009.

[260] Zhang, J., He, S., Ling, C. X., et al. PeakSelect: preprocessing tandem mass spectra for better peptide identification. *Rapid Communications in Mass Spectrometry*, 22(8):1203–1212, 2008.

[261] Eng, J. K., McCormack, A. L., and Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, November 1994.

[262] Perkins, D. N., Pappin, D. J. C., Creasy, D. M., et al. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, December 1999.

[263] Deutsch, E. W., Lam, H., and Aebersold, R. Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiological Genomics*, 33(1):18–25, March 2008.

[264] Nesvizhskii, A. I., Keller, A., Kolker, E., et al. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry. *Analytical Chemistry*, 75(17):4646–4658, September 2003.

[265] Li, X.-j., Zhang, H., Ranish, J. A., et al. Automated Statistical Analysis of Protein Abundance Ratios from Data Generated by Stable-Isotope Dilution and Tandem Mass Spectrometry. *Analytical Chemistry*, 75(23):6648–6657, October 2003.

[266] Cox, J. and Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12):1367–1372, November 2008.

[267] Rigden, D. J. and Fernández, X. M. The 2018 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Research*, 46(D1):D1–D7, January 2018.

[268] O'Leary, N. A., Wright, M. W., Brister, J. R., et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, January 2016.

[269] Frankish, A., Diekhans, M., Ferreira, A.-M., et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 306(Suppl. 1):636, October 2018.

[270] Brown, G. R., Hem, V., Katz, K. S., et al. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Research*, 43(D1):D36–D42, January 2015.

[271] Stelzer, G., Rosen, N., Plaschkes, I., et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics*, 54(1):1.30.1–1.30.33, June 2016.

[272] Yates, B., Braschi, B., Gray, K. A., et al. Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Research*, 45(D1):D619–D625, January 2017.

[273] Zerbino, D. R., Achuthan, P., Akanni, W., et al. Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761, January 2018.

[274] Bateman, A., Martin, M. J., O'Donovan, C., et al. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, January 2017.

[275] Matys, V., Kel-Margoulis, O. V., Fricke, E., et al. TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, 34:D108–D110, January 2006.

[276] Khan, A., Fornes, O., Stigliani, A., et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research*, 46(D1):D260–D266, January 2018.

[277] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012.

[278] Kozomara, A. and Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, 39(Database issue):D152–157, January 2011.

[279] Backes, C., Fehlmann, T., Kern, F., et al. miRCarta: a central repository for collecting miRNA candidates. *Nucleic Acids Research*, 46(D1):D160–D167, January 2018.

[280] Dweep, H., Sticht, C., Pandey, P., et al. miRWalk – Database: Prediction of possible miRNA binding sites by "walking" the genes of three genomes. *Journal of Biomedical Informatics*, 44(5):839–847, October 2011.

[281] The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research*, 45(D1):D331–D338, January 2017.

[282] Kanehisa, M., Sato, Y., Kawashima, M., et al. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462, January 2016.

[283] Fabregat, A., Jupe, S., Matthews, L., et al. The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, 46(D1):D649–D655, January 2018.

[284] Slenter, D. N., Kutmon, M., Hanspers, K., et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, 46(D1):D661–D667, January 2018.

[285] Kitts, A., Phan, L., Ward, M., et al. The Database of Short Genetic Variation (dbSNP). In *The NCBI Handbook*. National Center for Biotechnology Information (US), April 2014.

[286] Rubio-Perez, C., Tamborero, D., Schroeder, M. P., et al. In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities. *Cancer Cell*, 27(3):382–396, March 2015.

[287] Chung, I. F., Chen, C.-Y., Su, S.-C., et al. DriverDBv2: a database for human cancer driver gene research. *Nucleic Acids Research*, 44(D1):D975–D979, January 2016.

[288] Wishart, D. S., Feunang, Y. D., Guo, A. C., et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, January 2018.

[289] Cotto, K. C., Wagner, A. H., Feng, Y.-Y., et al. DGIdb 3.0: a redesign and expansion of the drug–gene interaction database. *Nucleic Acids Research*, 46(D1):D1068–D1073, November 2017.

[290] Li, Y. H., Yu, C. Y., Li, X. X., et al. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Research*, 46(D1):D1121–D1127, January 2018.

[291] Abramson, R. G. Overview of Targeted Therapies for Cancer. https://www.mycancergenome.org/content/page/overview-of-targeted-therapies-for-cancer/, May 2018.

[292] Chakravarty, D., Gao, J., Phillips, S., et al. OncoKB: A Precision Oncology Knowledge Base. *JCO Precision Oncology*, (1):1–16, July 2017.

[293] Yang, W., Soares, J., Greninger, P., et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(D1):D955–D961, January 2013.

[294] Barrett, T., Wilhite, S. E., Ledoux, P., et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995, January 2013.

[295] Kolesnikov, N., Hastings, E., Keays, M., et al. ArrayExpress update—simplifying data submissions. *Nucleic Acids Research*, 43(D1):D1113–D1116, October 2014.

[296] Leinonen, R., Sugawara, H., Shumway, M., et al. The Sequence Read Archive. *Nucleic Acids Research*, 39(Database):D19–D21, December 2010.

[297] Lappalainen, I., Almeida-King, J., Kumanduri, V., et al. The European Genome-phenome Archive of human data consented for biomedical research. *Nature Genetics*, 47(7):692–695, July 2015.

[298] Network, T. R. The Cancer Genome Atlas: Charting a New Course for Cancer Prevention, Diagnosis and Treatment. http://cancergenome.nih.gov/pdfs/TCGA_Program_Brochure_2014.

[299] Costello, J. C., Heiser, L. M., Georgii, E., et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology*, 32(12):1202–1212, December 2014.

[300] Fisher, R. A. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1958.

[301] Neyman, J. and Pearson, E. S. On the Problem of the Most Efficient Tests of Statistical Hypotheses. In *Breakthroughs in Statistics*, pages 73–108. Springer, New York, NY, 1992.

[302] Rohatgi, V. K. and Saleh, A. K. M. E. *An Introduction to Probability and Statistics*. John Wiley & Sons, second edition, 2001.

[303] Sullivan, G. M. and Feinn, R. Using Effect Size—or Why the PValue Is Not Enough. *Journal of Graduate Medical Education*, 4(3):279–282, September 2012.

[304] Kuffner, T. A. and Walker, S. G. Why are p-Values Controversial? *The American Statistician*, 73(1):1–3, January 2018.

[305] Rosenthal, R. and Rubin, D. B. The Counternull Value of an Effect Size: A New Statistic. *Psychological Science*, 5:329–334, November 1994.

[306] Bonferroni, C. E. Il calcolo delle assicurazioni su gruppi di teste. January 1935.

[307] Sidak, Z. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *Journal of the American Statistical Association*, 62(318):626, June 1967.

[308] Holm, S. and Sture. A Simple Sequentially Rejective Multiple Test Procedure. *Psychological Science*, 6:65–70, 1979.

[309] Finner, H. On a Monotonicity Problem in Step-Down Multiple Test Procedures. *Journal of the American Statistical Association*, 88(423):920–923, September 1993.

[310] Hochberg, Y. and Benjamini, Y. More powerful procedures for multiple significance testing. *Statistics in Medicine*, 9(7):811–818, July 1990.

[311] Benjamini, Y. and Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, August 2001.

[312] Kanji, G. K. *100 Statistical Tests*. SAGE Publishing, third edition, November 2018.

[313] Ruxton, G. D. The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. *Behavioral Ecology*, 17(4):688–690, July 2006.

[314] Zar, J. H. *Biostatistical Analysis*. Pearson, fifth edition, 2010.

[315] Student. The Probable Error of a Mean. *Biometrika*, 6(1):1, March 1908.

[316] Welch, B. L. The Significance of the Difference Between Two Means when the Population Variances are Unequal. *Biometrika*, 29(3/4):350, February 1938.

[317] Welch, B. L. The Generalization of 'Student's' Problem when Several Different Population Variances are Involved. *Biometrika*, 34(1/2):28, January 1947.

[318] Opgen-Rhein, R. and Strimmer, K. Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Statistical Applications in Genetics and Molecular Biology*, 6, 2007.

[319] Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80, December 1945.

[320] Bellera, C. A., Julien, M., and Hanley, J. A. Normal Approximations to the Distributions of the Wilcoxon Statistics: Accurate to What N? Graphical Insights. *Journal of Statistics Education*, 18(2), August 2017.

[321] Hartwell, L. H., Hopfield, J. J., Leibler, S., et al. From molecular to modular cell biology. *Nature*, 402(6761supp):C47–C52, December 1999.

[322] García-Campos, M. A., Espinal-Enríquez, J., and Hernández-Lemus, E. Pathway Analysis: State of the Art. *Frontiers in Physiology*, 6(278):47, December 2015.

[323] Khatri, P., Sirota, M., and Butte, A. J. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS computational biology*, 8(2):e1002375, February 2012.

[324] Casella, G. and Berger, R. *Statistical Inference*. Cengage Learning, second edition, July 2001.

[325] Backes, C., Keller, A., Kuentzer, J., et al. GeneTrail–advanced gene set enrichment analysis. *Nucleic Acids Research*, 35(Web Server issue):W186–192, July 2007.

[326] Subramanian, A., Tamayo, P., Mootha, V. K., et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, October 2005.

[327] Keller, A., Backes, C., and Lenhof, H.-P. Computation of significance scores of unweighted Gene Set Enrichment Analyses. *BMC Bioinformatics*, 8(1):290, August 2007.

[328] Massey Jr, F. J. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46(253):68–78, April 2012.

[329] Powers, R. K., Goodspeed, A., Pielke-Lombardo, H., et al. GSEA-InContext: identifying novel and common patterns in expression experiments. *Bioinformatics*, 34(13):i555–i564, July 2018.

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

[331] Efron, B. and Tibshirani, R. On Testing the Significance of Sets of Genes. *Psychological Science*, 1(1):107–129, July 2007.

[332] Ackermann, M. and Strimmer, K. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10(1):47, December 2009.

[333] Drier, Y., Sheffer, M., and Domany, E. Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 110(16):6388–6393, April 2013.

[334] Hastie, T. and Stuetzle, W. Principal Curves. *Journal of the American Statistical Association*, 84(406):502–516, March 2012.

[335] Livshits, A., Git, A., Fuks, G., et al. Pathway-based personalized analysis of breast cancer expression data. *Molecular Oncology*, 9(7):1471–1483, August 2015.

[336] Guo, Z., Zhang, T., Li, X., et al. Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics*, 6(1):58, 2005.

[337] Bild, A. H., Yao, G., Chang, J. T., et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, 439(7074):353–357, January 2006.

[338] Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[339] Lee, E., Chuang, H.-Y., Kim, J.-W., et al. Inferring Pathway Activity toward Precise Disease Classification. *PLoS computational biology*, 4(11):e1000217, July 2008.

[340] Sootanan, P., Prom-on, S., Meechai, A., et al. Microarray-Based Disease Classification Using Pathway Activities with Negatively Correlated Feature Sets. In *Neural Information Processing. Models and Applications*, pages 250–258. Springer, Berlin, Heidelberg, November 2010.

[341] Helms, V. *Principles of Computational Cell Biology*. From Protein Complexes to Cellular Networks. Wiley, 2008.

[342] Rahnenführer, J., Domingues, F. S., Maydt, J., et al. Calculating the Statistical Significance of Changes in Pathway Activity From Gene Expression Data. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–29, June 2004.

[343] Pearson, K. VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242, January 1895.

[344] Papoulis, A. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, third edition, 1991.

[345] Dangeti, P. *Statistics for Machine Learning*. Packt Publishing, July 2017.

[346] Tarca, A. L., Draghici, S., Khatri, P., et al. A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82, January 2009.

[347] Brin, S. and Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. http://ilpubs.stanford.edu:8090/361/, 1998.

[348] Bokanizad, B., Tagett, R., Ansari, S., et al. SPATIAL: A System-level PAThway Impact AnaLysis approach. *Nucleic Acids Research*, May 2016.

[349] Vandin, F., Upfal, E., and Raphael, B. J. Algorithms for Detecting Significantly Mutated Pathways in Cancer. *Journal of Computational Biology*, 18(3):507–522, March 2011.

[350] Feller, W. *An Introduction to Probability Theory and Its Applications - Volume II*. John Wiley & Sons, 1970.

[351] Masuda, N., Porter, M. A., and Lambiotte, R. Random walks and diffusion on networks. *Physics Reports*, 716-717:1–58, November 2017.

[352] Glaab, E., Baudot, A., Krasnogor, N., et al. EnrichNet: network-based gene set enrichment analysis. *Bioinformatics*, 28(18):i451–i457, September 2012.

[353] Paull, E. O., Carlin, D. E., Niepel, M., et al. Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics*, 29(21):2757–2764, November 2013.

[354] Dimitrakopoulos, C., Hindupur, S. K., Häfliger, L., et al. Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics*, 34(14):2441–2448, March 2018.

[355] Keller, A., Backes, C., Gerasch, A., et al. A novel algorithm for detecting differentially regulated paths based on gene set enrichment analysis. *Bioinformatics*, 25(21):2787–2794, November 2009.

[356] Ideker, T., Ozier, O., Schwikowski, B., et al. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18:S233–S240, July 2002.

[357] Ulitsky, I., Krishnamurthy, A., Karp, R. M., et al. DEGAS: De Novo Discovery of Dysregulated Pathways in Human Diseases. *PLOS ONE*, 5(10), October 2010.

[358] Alcaraz, N., Kücük, H., Weile, J., et al. KeyPathwayMiner: Detecting Case-Specific Biological Pathways Using Expression Data. *Internet Mathematics*, 7(4):299–313, November 2011.

[359] Dittrich, M. T., Klau, G. W., Rosenwald, A., et al. Identifying functional modules in protein–protein interaction networks: an integrated exact approach. *Bioinformatics*, 24(13):i223–i231, July 2008.

[360] Ljubic, I., Weiskircher, R., Pferschy, U., et al. Solving the Prize-Collecting Steiner Tree Problem to Optimality. *ALENEX/ANALCO*, pages 68–76, 2005.

[361] Mehlhorn, K. and Sanders, P. *Algorithms and Data Structures - The Basic Toolbox*. Springer, 2008.

[362] Backes, C., Rurainski, A., Klau, G. W., et al. An integer linear programming approach for finding deregulated subgraphs in regulatory networks. *Nucleic Acids Research*, 40(6):e43, March 2012.

[363] Schwartz, A. S., Yu, J., Gardenour, K. R., et al. Cost-effective strategies for completing the interactome. *Nature Methods*, 6(1):55–61, January 2009.

[364] Deane, C. M., Salwiński, Ł., Xenarios, I., et al. Protein Interactions: Two Methods for Assessment of the Reliability of High Throughput Observations. *Molecular & Cellular Proteomics*, 1(5):349–356, May 2002.

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

[366] Kanehisa, M. and Bork, P. Bioinformatics in the post-sequence era. *Nature Genetics*, 33(3s):305–310, March 2003.

[367] Ghosh, S., Matsuoka, Y., Asai, Y., et al. Software for systems biology: from tools to integrated platforms. *Nature Reviews Genetics*, 12(12):821–832, December 2011.

[368] Association, S. I. I. Software as a Service - Strategic Backgrounder. https://www.edocr.com/v/83ev4lkw/edocr/Software-as-a-Service-Strategic-Backgrounder. Technical report, February 2001.

[369] Kumar, S. and Dudley, J. Bioinformatics software for biologists in the genomics era. *Bioinformatics*, 23(14):1713–1717, July 2007.

[370] Swertz, M. A. and Jansen, R. C. Beyond standardization: dynamic software infrastructures for systems biology. *Nature Reviews Genetics*, 8(3):235–243, February 2007.

[371] Berthold, M. R., Cebron, N., Dill, F., et al. KNIME - the Konstanz information miner: version 2.0 and beyond. *ACM SIGKDD Explorations Newsletter*, 11(1):26–31, November 2009.

[372] Wolstencroft, K., Haines, R., Fellows, D., et al. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Research*, 41(Web Server issue):W557–W561, July 2013.

[373] Blankenberg, D., Von Kuster, G., Coraor, N., et al. Galaxy: A Web-Based Genome Analysis Tool for Experimentalists. *Current Protocols in Molecular Biology*, 89(1):19.10.1–19.10.21, January 2010.

[374] Faulkner, S., Eicholz, A., Leithead, T., et al. HTML 5.2. https://www.w3.org/TR/html52/. Technical report, December 2017.

[375] Atkins Jr, T., Etemad, E. J., and Rivoal, F. CSS Snapshot 2017. https://www.w3.org/TR/css-2017/. Technical report, January 2017.

[376] Otto, Marc, Thornton, Jacob, Rebert, Chris, et al. Bootstrap. https://github.com/twbs/bootstrap/tree/v4.3.1.

[377] Fernandez, D. Thymeleaf template engine. http://www.thymeleaf.org/. Technical report, 2015.

[378] Jardine, A. DataTables. https://datatables.net.

[379] Timberg, E., Linsley, T., Brunel, S., et al. Charts.js. https://www.chartjs.org.

[380] Bostock, M., Ogievetsky, V., and Heer, J. D 3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, December 2011.

[381] Highsoft. Highcharts. https://www.highcharts.com.

[382] Arnold, K., Gosling, J., and Holmes, D. *The Java Programming Language*. Addison Wesley Professional, fourth edition, August 2005.

[383] Flanagan, D. *JavaScript - The Definitive Guide*. O'Reilly Media, Inc, fifth edition, August 2006.

[384] Duckett, J. *JavaScript and JQuery*. Interactive Front-End Web Development. Wiley Publishing, first edition, 2014.

[385] Eichorn, J. *Understanding AJAX - Using JavaScript to Create Rich Internet Applications*. Prentice Hall PTR, 2006.

[386] Massé, M. *Rest API Design Rulebook - Designing Consistent RESTful Web Service Interfaces*. O'Reilly Media, Inc., 2011.

[387] Lippman, S. B., Lajoie, J., and Moo, B. E. *C++ Primer*. Addison-Wesely, fourth edition, July 2008.

[388] Worsley, John C and Drake, Joshua D. Random walks and diffusion on networks. *Physics Reports*, 716-717:1–58, November 2017.

[389] Moniruzzaman, A. B. M. and Hossain, S. A. NoSQL Database: New Era of Databases for Big data Analytics - Classification, Characteristics and Comparison. *International Journal of Database Theory and Application*, 6(4), June 2013.

[390] Fielding, R. T. *Architectural Styles and the Design of Network-based Software Architectures*. Ph.D. thesis, 2000.

[391] Tusher, V. G., Tibshirani, R., and Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5116–5121, April 2001.

[392] Larsen, R. J. and Marx, M. L. *Introduction to Mathematical Statistics and Its Applications: Pearson New International Edition*. Pearson Education, Limited, July 2013.

[393] Welch, B. L. The Generalization of 'Student's' Problem when Several Different Population Variances are Involved. *Biometrika*, 34(1/2):28–35, January 1947.

[394] Student. The Probable Error of a Mean. *Biometrika*, 6:1–25, 1908.

[395] Mann, H. B. and Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60, March 1947.

[396] Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–442, January 2002.

[397] William H. Press, Teukolsky, S. A., Vetterling, W. T., et al. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, third edition, September 2007.

[398] Lander, E. S., Linton, L. M., Birren, B., et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[399] Harrow, J., Frankish, A., Gonzalez, J. M., et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*, 22(9):1760–1774, September 2012.

[400] Stöckel, D. *Bioinformatics methods for the genetic and molecular characterisation of cancer*. Ph.D. thesis, 2016.

[401] Backes, C., Keller, A., Kuentzer, J., et al. GeneTrail–advanced gene set enrichment analysis. *Nucleic Acids Research*, 35(Web Server):W186–W192, May 2007.

[402] van de Vijver, M. J., He, Y. D., van 't Veer, L. J., et al. A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. *New England Journal of Medicine*, 347(25):1999–2009, December 2002.

[403] Fielden, M. R., Brennan, R., and Gollub, J. A Gene Expression Biomarker Provides Early Prediction and Mechanistic Assessment of Hepatic Tumor Induction by Nongenotoxic Chemicals. *Toxicological Sciences*, 99(1):90–100, September 2007.

[404] Garcia-Aguilar, J., Chen, Z., Smith, D. D., et al. Identification of a biomarker profile associated with resistance to neoadjuvant chemoradiation therapy in rectal cancer. *Annals of surgery*, 254(3):486–493, September 2011.

[405] Drabovich, A. P., Saraon, P., Drabovich, M., et al. Multi-omics biomarker pipeline reveals elevated levels of protein-glutamine gamma-glutamyltransferase 4 in seminal plasma of prostate cancer patients. *Molecular & Cellular Proteomics*, 18(9):1807–1823, May 2018.

[406] Ein-Dor, L., Kela, I., Getz, G., et al. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178, January 2005.

[407] Lee, E., Chuang, H.-Y., Kim, J.-W., et al. Inferring pathway activity toward precise disease classification. *PLoS computational biology*, 4(11):e1000217, November 2008.

[408] Doniger, S. W., Salomonis, N., Dahlquist, K. D., et al. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome biology*, 4(1):R7, 2003.

[409] Draghici, S., Khatri, P., Martins, R. P., et al. Global functional profiling of gene expression. *Science*, 81(2):98–104, February 2003.

[410] Pavlidis, P., Qin, J., Arango, V., et al. Using the Gene Ontology for Microarray Data Mining: A Comparison of Methods and Application to Age Effects in Human Prefrontal Cortex. *Neurochemical Research*, 29(6):1213–1222.

[411] Tian, L., Greenberg, S. A., Kong, S. W., et al. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38):13544–13549, September 2005.

[412] Chuang, H.-Y., Lee, E., Liu, Y.-T., et al. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3:D418, 2007.

[413] Huang, D. W., Sherman, B. T., and Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, January 2009.

[414] Zeeberg, B. R., Feng, W., Wang, G., et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome biology*, 4(4):R28, April 2003.

[415] Subramanian, A., Kuehn, H., Gould, J., et al. GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics*, 23(23):3251–3253, December 2007.

[416] Subramanian, A., Tamayo, P., Mootha, V. K., et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, October 2005.

[417] Alonso, R., Salavert, F., Garcia-Garcia, F., et al. Babelomics 5.0: functional interpretation for new generations of genomic data. *Nucleic Acids Research*, 43(W1):W117–W121, June 2015.

[418] Eden, E., Navon, R., Steinfeld, I., et al. GOrilla : a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10(1):48, December 2009.

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

[420] Henry, V. J., Bandrowski, A. E., Pepin, A.-S., et al. OMICtools: an informative directory for multi-omic data analysis. *Database: The Journal of Biological Databases and Curation*, 2014(0):bau069–bau069, 2014.

[421] Huang, D. W., Sherman, B. T., and Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, January 2009.

[422] Hung, J.-H., Yang, T.-H., Hu, Z., et al. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in Bioinformatics*, 13(3):281–291, May 2012.

[423] Naeem, H., Zimmer, R., Tavakkolkhah, P., et al. Rigorous assessment of gene set enrichment tests. *Bioinformatics*, 28(11):1480–1486, June 2012.

[424] Ashburner, M., Ball, C. A., Blake, J. A., et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29, May 2000.

[425] Croft, D., Mundo, A. F., Haw, R., et al. The Reactome pathway knowledgebase. *Nucleic Acids Research*, page gkt1102, November 2013.

[426] Kelder, T., van Iersel, M. P., Hanspers, K., et al. WikiPathways: building research communities on biological pathways. *Nucleic Acids Research*, 40(Database issue):D1301–D1307, January 2012.

[427] Law, V., Knox, C., Djoumbou, Y., et al. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research*, 42(Database issue):D1091–1097, January 2014.

[428] Wong, N. and Wang, X. miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Research*, 43(Database issue):D146–152, January 2015.

[429] Zimmer, B. and Kerren, A. OnGraX: A Web-Based System for the Collaborative Visual Analysis of Graphs. *Journal of Graph Algorithms and Applications*, 21(1):5–27, 2017.

[430] Pancreatic Cancer - Statistics. *https://www.cancer.net/cancer-types/pancreatic-cancer/statistics*, January 2019.

[431] Becker, A. E., Hernandez, Y. G., Frucht, H., et al. Pancreatic ductal adenocarcinoma: Risk factors, screening, and early detection. *World Journal of Gastroenterology : WJG*, 20(32):11182, August 2014.

[432] Ryan, D. P., Hong, T. S., and Bardeesy, N. Pancreatic Adenocarcinoma. *New England Journal of Medicine*, 371(11):1039–1049, September 2014.

[433] Adamska, A., Domenichini, A., and Falasca, M. Pancreatic Ductal Adenocarcinoma: Current and Evolving Therapies. *International Journal of Molecular Sciences*, 18(7):1338, July 2017.

[434] Raphael, B. J., Hruban, R. H., Aguirre, A. J., et al. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer Cell*, 32(2):185–203.e13, August 2017.

[435] Bailey, P., Chang, D. K., Nones, K., et al. Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature*, 531(7592):47–52, March 2016.

[436] Collisson, E. A., Sadanandam, A., Olson, P., et al. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nature Medicine*, 17(4):500–503, April 2011.

[437] Kalkat, M., De Melo, J., Hickman, K., et al. MYC Deregulation in Primary Human Cancers. *Genes*, 8(6):151, June 2017.

[438] Schaub, F. X., Dhankani, V., Trivedi, M., et al. Pan-cancer Alterations of the MYC Oncogene and Its Proximal Network across the Cancer Genome Atlas. *Cell Systems*, 6(3):282–300.e2, March 2018.

[439] Wirth, M., Mahboobi, S., Krämer, O. H., et al. Concepts to Target MYC in Pancreatic Cancer. *Molecular Cancer Therapeutics*, 15(8):1792–1798, August 2016.

[440] Chen, H., Liu, H., and Qing, G. Targeting oncogenic Myc as a strategy for cancer treatment. *Signal Transduction and Targeted Therapy*, 3(1):5, February 2018.

[441] Dang, C. V. MYC on the Path to Cancer. *Cell*, 149(1):22–35, March 2012.

[442] Miller, D. M., Thomas, S. D., Islam, A., et al. c-Myc and Cancer Metabolism. *Clinical Cancer Research*, 18(20):5546–5553, October 2012.

[443] Beijersbergen, R. L., Wessels, L. F. A., and Bernards, R. Synthetic Lethality in Cancer Therapeutics. *Annual Review of Cancer Biology*, 1(1):141–161, March 2017.

[444] O'Neil, N. J., Bailey, M. L., and Hieter, P. Synthetic lethality and cancer. *Nature Reviews Genetics*, 18(10):613–623, October 2017.

[445] Cermelli, S., Jang, I. S., Bernard, B., et al. Synthetic Lethal Screens as a Means to Understand and Treat MYC-Driven Cancers. *Cold Spring Harbor Perspectives in Medicine*, 4(3), March 2014.

[446] Kessler, J. D., Kahle, K. T., Sun, T., et al. A SUMOylation-Dependent Transcriptional Subprogram Is Required for Myc-Driven Tumorigenesis. *Science*, 335(6066):348–353, January 2012.

[447] Hoellein, A., Fallahi, M., Schoeffmann, S., et al. Myc-induced SUMOylation is a therapeutic vulnerability for B-cell lymphoma. *Blood*, 124(13):2081–2090, September 2014.

[448] Hickey, C. M., Wilson, N. R., and Hochstrasser, M. Function and regulation of SUMO proteases. *Nature Reviews Molecular Cell Biology*, 13(12):755–766, December 2012.

[449] Liberzon, A., Birger, C., Thorvaldsdóttir, H., et al. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, 1(6):417–425, December 2015.

[450] He, X., Riceberg, J., Soucy, T., et al. Probing the roles of SUMOylation in cancer cell biology by using a selective SAE inhibitor. *Nature Chemical Biology*, 13(11):1164–1171, September 2017.

[451] Lee, T. I. and Young, R. A. Transcriptional Regulation and Its Misregulation in Disease. *Cell*, 152(6):1237–1251, March 2013.

[452] Nebert, D. W. Transcription factors and cancer: an overview. *Toxicology*, 181-182:131–141, December 2002.

[453] Patricia A J Muller, K. H. V. Mutant p53 in Cancer: New Functions and Therapeutic Opportunities. *Cancer Cell*, 25(3):304–317, March 2014.

[454] Darnell, J. E. Transcription factors as targets for cancer therapy. *Nature Reviews Cancer*, 2(10):740–749, October 2002.

[455] Anand S Bhagwat, C. R. V. Targeting Transcription Factors in Cancer. *Trends in cancer*, 1(1):53–65, September 2015.

[456] Essaghir, A., Toffalini, F., Knoops, L., et al. Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. *Nucleic Acids Research*, 38(11), June 2010.

[457] Yang, J., Yu, H., Liu, B.-H., et al. DCGL v2.0: An R Package for Unveiling Differential Regulation from Differential Co-expression. *PLOS ONE*, 8(11), November 2013.

[458] Liu, B.-H., Yu, H., Tu, K., et al. DCGL: an R package for identifying differentially coexpressed genes and links from gene expression microarray data . *Biometrika*, 26(20), August 2010.

[459] Reverter, A., Hudson, N. J., Nagaraj, S. H., et al. Regulatory impact factors: unraveling the transcriptional regulation of complex traits from expression data. *Bioinformatics*, 26(7):896–904, April 2010.

[460] Huang, C.-L., Lamb, J., Chindelevitch, L., et al. Correlation set analysis: detecting active regulators in disease populations using prior causal knowledge . *BMC Bioinformatics*, 13(1):46, December 2012.

[461] Gonçalves, J. P., Francisco, A. P., Mira, N. P., et al. TFRank: network-based prioritization of regulatory associations underlying transcriptional responses. *Bioinformatics*, 27(22):3149–3157, November 2011.

[462] Poos, A. M., Maicher, A., Dieckmann, A. K., et al. Mixed Integer Linear Programming based machine learning approach identifies regulators of telomerase in yeast. *Nucleic Acids Research*, 44(10), June 2016.

[463] Gonçalves, J. P., Aires, R. S., Francisco, A. P., et al. Regulatory Snapshots: Integrative Mining of Regulatory Modules from Expression Time Series and Regulatory Networks. *PLOS ONE*, 7(5), January 2012.

[464] Pique-Regi, R., Degner, J. F., Pai, A. A., et al. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, 21(3):447–455, March 2011.

[465] Luo, K. and Hartemink, A. J. Using DNAse digestion data to accurately identify transcription factor binding sites. In *Proceedings of the Pacific Symposium on Biocomputing 2013*, pages 80–91. WORLD SCIENTIFIC, November 2012.

[466] Schmidt, F., Gasparoni, N., Gasparoni, G., et al. Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Research*, 45(1):54–66, January 2017.

[467] Schmidt, F., Kern, F., Ebert, P., et al. TEPIC 2—an extended framework for transcription factor binding prediction and integrative epigenomic analysis. *Bioinformatics*, 28:56, October 2018.

[468] Lachmann, A., Xu, H., Krishnan, J., et al. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*, 26(19):2438–2444, October 2010.

[469] Yang, J.-H., Li, J.-H., Jiang, S., et al. ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data. *Nucleic Acids Research*, 41(D1):D177–D187, January 2013.

[470] Sloan, C. A., Chan, E. T., Davidson, J. M., et al. ENCODE data at the ENCODE portal. *Nucleic Acids Research*, 44(D1):D726–D732, January 2016.

[471] Mathelier, A., Fornes, O., Arenillas, D. J., et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 44(D1):D110–D115, January 2016.

[472] Fazekas, D., Koltai, M., Türei, D., et al. SignaLink 2 – a signaling pathway resource with multi-layered regulatory networks. *BMC systems biology*, 7(1):7, December 2013.

[473] Bailey, T. L. Discovering Sequence Motifs. In *Bioinformatics*, pages 231–251. Humana Press, 2008.

[474] Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., et al. HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Research*, 44(D1):D116–D125, January 2016.

[475] Kheradpour, P. and Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Research*, 42(5):2976–2987, March 2014.

[476] Song, L. and Crawford, G. E. DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells. *Cold Spring Harbor Protocols*, 2010(2), February 2010.

[477] Niwa, H. Open conformation chromatin and pluripotency. *Genes & Development*, 21(21):2671–2676, November 2007.

[478] Hoerl, A. E. and Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 42(1):80, 1970.

[479] Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Psychological Science*, 58(1):267–288, 1996.

[480] Zou, H. and Hastie, T. Regularization and Variable Selection via the Elastic Net. *Psychological Science*, 67(2):301–320, 2005.

[481] Spearman, C. The proof and measurement of association between two things. *International Journal of Epidemiology*, 39(5):1137–1150, 1904.

[482] Hastie, T., Tibshirani, R., and Friedman, J. H. *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. Springer, second edition, 2009.

[483] Pearson, K. Notes on regression and inheritance in the case of two parents. In *Proceedings of the Royal Society of London*. June 1895.

[484] Efron, B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26, January 1979.

[485] Kehl, T., Schneider, L., Kattler, K., et al. REGGAE: a novel approach for the identification of key transcriptional regulators. *Bioinformatics*, 34(20):3503–3510, October 2018.

[486] Siegel, R. L., Miller, K. D., and Jemal, A. Cancer statistics, 2017. *CA: A Cancer Journal for Clinicians*, 67(1):7–30, January 2017.

[487] Fillmore, C. M., Gupta, P. B., Rudnick, J. A., et al. Estrogen expands breast cancer stem-like cells through paracrine FGF/Tbx3 signaling. *Proceedings of the National Academy of Sciences of the United States of America*, 107(50):21737–21742, December 2010.

[488] Bae, S. Y., Kim, S., Lee, J. H., et al. Poor prognosis of single hormone receptor- positive breast cancer: similar outcome as triple-negative breast cancer. *BMC cancer*, 15(1):138, December 2015.

[489] Heiser, L. M., Sadanandam, A., Kuo, W.-L., et al. Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 109(8):2724–2729, February 2012.

[490] Neve, R. M., Chin, K., Fridlyand, J., et al. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell*, 10(6):515–527, December 2006.

[491] Mehra, R., Varambally, S., Ding, L., et al. Identification of GATA3 as a Breast Cancer Prognostic Marker by Global Gene Expression Meta-analysis. *Cancer Research*, 65(24):11259–11264, December 2005.

[492] Mehta, R. J., Jain, R. K., Leung, S., et al. FOXA1 is an independent prognostic marker for ER-positive breast cancer. *Breast Cancer Research and Treatment*, 131(3):881–890, April 2011.

[493] van 't Veer, L. J., Dai, H., van de Vijver, M. J., et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, January 2002.

[494] West, M., Blanchette, C., Dressman, H., et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 98(20):11462–11467, September 2001.

[495] Kong, S. L., Li, G., Loh, S. L., et al. Cellular reprogramming by the conjoint action of ERα, FOXA1, and GATA3 to a ligand-inducible growth state. *Molecular Systems Biology*, 7(1):526–526, January 2011.

[496] Sachs, M., Onodera, C., Blaschke, K., et al. Bivalent Chromatin Marks Developmental Regulatory Genes in the Mouse Embryonic Germline In Vivo. *Cell Reports*, 3(6):1777–1784, June 2013.

[497] Fletcher, M. N. C., Castro, M. A. A., Wang, X., et al. Master regulators of FGFR2 signalling and breast cancer risk. *Nature Communications*, 4(1):392, September 2013.

[498] Davidoff, A. M. Wilms Tumor. *Advances in Pediatrics*, 59(1):247–267, January 2012.

[499] Furtwängler, R., Nourkami, N., Alkassar, M., et al. Update on Relapses in Unilateral Nephroblastoma Registered in 3 Consecutive SIOP/GPOH Studies – A Report from the GPOH-Nephroblastoma Study Group. *Klinische Pädiatrie*, 223(03):113–119, April 2011.

[500] Popov, S. D., Sebire, N. J., and Vujanic, G. M. Wilms' Tumour – Histology and Differential Diagnosis. In *Wilms Tumor*, pages 3–21. Codon Publications, March 2016.

[501] van den Heuvel-Eibrink, M. M., van Tinteren, H., Bergeron, C., et al. Outcome of localised blastemal-type Wilms tumour patients treated according to intensified treatment in the SIOP WT 2001 protocol, a report of the SIOP Renal Tumour Study Group (SIOP-RTSG). *European Journal of Cancer*, 51(4):498–506, March 2015.

[502] Rosa, A. and Brivanlou, A. H. A regulatory circuitry comprised of miR-302 and the transcription factors OCT4 and NR2F2 regulates human embryonic stem cell differentiation. *The EMBO journal*, 30(2):237–248, January 2011.

[503] Qin, J., Chen, X., Xie, X., et al. COUP-TFII regulates tumor growth and metastasis by modulating tumor angiogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 107(8):3687–3692, February 2010.

[504] Slyper, M., Shahar, A., Bar-Ziv, A., et al. Control of Breast Cancer Growth and Initiation by the Stem Cell–Associated Transcription Factor TCF3. *Cancer Research*, 72(21):5613–5624, November 2012.

[505] Stöckel, D., Kehl, T., Trampert, P., et al. Multi-omics enrichment analysis using the GeneTrail2 web service. *Bioinformatics*, 32(10):1502–1508, May 2016.

[506] Losada, A. Cohesin in cancer: chromosome segregation and beyond. *Nature Reviews Cancer*, 14(6):389–393, June 2014.

[507] Noda, T., Nagano, H., Takemasa, I., et al. Activation of Wnt/β-catenin signalling pathway induces chemoresistance to interferon-α/5-fluorouracil combination therapy for hepatocellular carcinoma. *British Journal of Cancer*, 100(10):1647–1658, April 2009.

[508] Wang, L., Brugge, J. S., and Janes, K. A. Intersection of FOXO- and RUNX1-mediated gene expression programs in single breast epithelial cells during morphogenesis and tumor progression. *Proceedings of the National Academy of Sciences of the United States of America*, 108(40):E803–E812, October 2011.

[509] Della Gatta, G., Palomero, T., Perez-Garcia, A., et al. Reverse engineering of TLX oncogenic transcriptional networks identifies RUNX1 as tumor suppressor in T-ALL. *Nature Medicine*, 18(3):436–440, February 2012.

[510] North, T. E., Stacy, T., Matheny, C. J., et al. Runx1 Is Expressed in Adult Mouse Hematopoietic Stem Cells and Differentiating Myeloid and Lymphoid Cells, But Not in Maturing Erythroid Cells. *Stem Cells*, 22(2):158–168, March 2004.

[511] Comino-Méndez, I., Leandro-García, L. J., Montoya, G., et al. Functional and in silico assessment of MAX variants of unknown significance. *Journal of Molecular Medicine*, 93(11):1247–1255, June 2015.

[512] Juan, D., Perner, J., Carrillo de Santa Pau, E., et al. Epigenomic Co-localization and Co-evolution Reveal a Key Role for 5hmC as a Communication Hub in the Chromatin Network of ESCs. *Cell Reports*, 14(5):1246–1257, February 2016.

[513] Franz, M., Lopes, C. T., Huck, G., et al. Cytoscape.js: a graph theory library for visualisation and analysis . *Biometrika*, 32(12):309–311, January 2016.

[514] Gerasch, A., Faber, D., Küntzer, J., et al. BiNA: A Visual Analytics Tool for Biological Network Data. *PLoS ONE*, 9(2), February 2014.

[515] Shannon, P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504, November 2003.

[516] Wolkenhauer, O., Auffray, C., Jaster, R., et al. The road from systems biology to systems medicine. *Pediatric Research*, 73(4-2):502–507, January 2013.

[517] Griffith, M., Griffith, O. L., Coffman, A. C., et al. DGIdb: mining the druggable genome. *Nature Methods*, 10(12):1209–1210, December 2013.

[518] Thorn, C. F., Klein, T. E., and Altman, R. B. PharmGKB: the Pharmacogenomics Knowledge Base. *Methods in Molecular Biology*, 1015:311–320, 2013.

[519] Shrager, J., Tenenbaum, J. M., and Travers, M. Cancer Commons: Biomedicine in the internet age. In *Collaborative Computational Technologies for Biomedical Research*. John Wiley & Sons, 2011.

[520] Kuhn, M., Szklarczyk, D., Pletscher-Frankild, S., et al. STITCH 4: integration of protein-chemical interactions with user data. *Nucleic Acids Research*, 42(D1):D401–D407, January 2014.

[521] Bulusu, K. C., Tym, J. E., Coker, E. A., et al. canSAR: updated cancer research and drug discovery knowledgebase. *Nucleic Acids Research*, 42(Database issue):D1040–1047, January 2014.

[522] Tanoli, Z., Alam, Z., Ianevski, A., et al. Interactive visual analysis of drug–target interaction networks using Drug Target Profiler, with applications to precision medicine and drug repurposing. *Briefings in Bioinformatics*, 2018.

[523] Tang, J., Tanoli, Z.-u.-R., Ravikumar, B., et al. Drug Target Commons: A Community Effort to Build a Consensus Knowledge Base for Drug-Target Interactions. *Cell Chemical Biology*, 25(2):224–229.e2, February 2018.

[524] Lamb, J., Crawford, E. D., Peck, D., et al. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science*, 313(5795):1929–1935, September 2006.

[525] Duan, Q., Flynn, C., Niepel, M., et al. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. *Nucleic Acids Research*, 42(Web Server issue):W449–460, July 2014.

[526] Zhong, Y., Chen, E. Y., Liu, R., et al. Renoprotective effect of combined inhibition of angiotensin-converting enzyme and histone deacetylase. *Journal of the American Society of Nephrology*, 24(5):801–811, April 2013.

[527] Vera-Licona, P., Bonnet, E., Barillot, E., et al. OCSANA: optimal combinations of interventions from network analysis. *Bioinformatics*, 29(12):1571–1573, June 2013.

[528] Iadevaia, S., Lu, Y., Morales, F. C., et al. Identification of optimal drug combinations targeting cellular networks: integrating phospho-proteomics and computational network analysis. *Cancer Research*, 70(17):6704–6714, September 2010.

[529] Stamatakos, G. S., Dionysiou, D. D., Graf, N. M., et al. The "Oncosimulator": a multilevel, clinically oriented simulation system of tumor growth and organism response to therapeutic schemes. Towards the clinical evaluation of in silico oncology. *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2007:6629–6632, 2007.

[530] Peifer, M., Weiss, J., Sos, M. L., et al. Analysis of Compound Synergy in High-Throughput Cellular Screens by Population-Based Lifetime Modeling. *PLOS ONE*, 5(1), January 2010.

[531] Foo, J. and Michor, F. Evolution of Resistance to Targeted Anti-Cancer Therapies during Continuous and Pulsed Administration Strategies. *PLoS computational biology*, 5(11), November 2009.

[532] Yang, W., Soares, J., Greninger, P., et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(D1):D955–D961, January 2013.

[533] Iorio, F., Knijnenburg, T. A., Vis, D. J., et al. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, 166(3):740–754, July 2016.

[534] Kumar, R., Chaudhary, K., Gupta, S., et al. CancerDR: cancer drug resistance database. *Scientific Reports*, 3:1445, 2013.

[535] Gönen, M. Bayesian Efficient Multiple Kernel Learning. In *Proceedings of the 29th International Conference on Machine Learning*. 2012.

[536] Aben, N., Vis, D. J., Michaut, M., et al. TANDEM: a two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. *Bioinformatics*, 32(17):i413–i420, September 2016.

[537] Knijnenburg, T. A., Klau, G. W., Iorio, F., et al. Logic models to predict continuous outputs based on binary inputs with an application to personalized cancer therapy. *Scientific Reports*, 6(1):36812, November 2016.

[538] Hsu, S.-D., Tseng, Y.-T., Shrestha, S., et al. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Research*, 42(Database issue):D78–85, January 2014.

[539] Wingender, E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Briefings in Bioinformatics*, 9(4):326–332, July 2008.

[540] Whirl-Carrillo, M., McDonagh, E. M., Hebert, J. M., et al. Pharmacogenomics knowledge for personalized medicine. *Clinical pharmacology and therapeutics*, 92(4):414–417, October 2012.

[541] American Cancer Society. Cancer Facts & Figures 2014. http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2014/. Technical report, 2014.

[542] Schmoll, H. J., Cutsem, E. V., Stein, A., et al. ESMO Consensus Guidelines for management of patients with colon and rectal cancer. A personalized approach to clinical decision making. *Annals of Oncology*, 23(10):2479–2516, October 2012.

[543] Deutsche Gesellschaft für Hämatololgie und Medizinische Onkologie. Leitlinien zur Diagnostik und Therapie von Blut- und Krebserkrankungen. www.dgho-onkopedia.de. Technical report.

[544] Barrett, T., Wilhite, S. E., Ledoux, P., et al. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Research*, 41(Database issue):D991–995, January 2013.

[545] Li, H., Handsaker, B., Wysoker, A., et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

[546] DePristo, M. A., Banks, E., Poplin, R., et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498, May 2011.

[547] Koboldt, D. C., Zhang, Q., Larson, D. E., et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576, March 2012.

[548] McLaren, W., Pritchard, B., Rios, D., et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, 26(16):2069–2070, August 2010.

[549] Rana, P. and Sridhar, S. S. Efficacy and tolerability of lapatinib in the management of breast cancer. *Breast Cancer: Basic and Clinical Research*, 6:67–77, 2012.

[550] IBM. IBM CPLEX Optimizer. http://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/. Technical report, December 2014.

[551] NCBI. PubMed. http://www.ncbi.nlm.nih.gov/pubmed. Technical report.

[552] Vujanić, G. M. and Sandstedt, B. The pathology of Wilms' tumour (nephroblastoma): the International Society of Paediatric Oncology approach. *Journal of Clinical Pathology*, 63(2):102–109, February 2010.

[553] SIOP. International Society of Paediatric Oncology. http://www.siop-online.org/. Technical report.

[554] Agilent. GeneSpring GX. http://www.genomics.agilent.com/en/Microarray-Data-Analysis-Software/GeneSpring-GX/?cid=AG-PT-130&tabId=AG-PR-1061. Technical report, 2015.

[555] Muscal, J. A., Thompson, P. A., Horton, T. M., et al. A Phase I Trial of Vorinostat and Bortezomib in Children with Refractory or Recurrent Solid Tumors: A Children's Oncology Group Phase I Consortium Study (ADVL0916). *Pediatric Blood & Cancer*, 60(3):390–395, March 2013.

[556] Ross, S. A., McCaffery, P. J., Drager, U. C., et al. Retinoids in Embryonal Development. *Physiological Reviews*, 80(3):1021–1054, July 2000.

[557] Zirn, B., Samans, B., Spangenberg, C., et al. All-trans retinoic acid treatment of Wilms tumor cells reverses expression of genes associated with high risk and relapse in vivo. *Oncogene*, 24(33):5246–5251, May 2005.

[558] Wegert, J., Bausenwein, S., Kneitz, S., et al. Retinoic acid pathway activity in wilms tumors and characterization of biological responses in vitro. *Molecular Cancer*, 10(1):136, November 2011.

[559] Freemantle, S. J., Spinella, M. J., and Dmitrovsky, E. Retinoids in cancer therapy and chemoprevention: promise meets resistance. *Oncogene*, 22(47):7305–7315, October 2003.

[560] German Guideline Program in Oncology. Evidenced-based Guideline for Colorectal Cancer. http://www.awmf.org/fileadmin/user_upload/Leitlinien/021_D_Ges_fuer_Verdauungs-_und_Stoffwechselkrankheiten/021-007_S3_Colorectal_Cancer_2015_03.pdf. Technical report.

[561] Kandoth, C. vcf2maf. https://github.com/ ckandoth/vcf2maf. Technical report.

[562] American Cancer Society. What are key statistics about lung cancer? http://www.cancer.org/cancer/lungcancer-non-smallcell/detailedguide/non-small-cell-lung-cancer-key-statistics. Technical report.

[563] Reck, M., Popat, S., Reinmuth, N., et al. Metastatic non-small cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 23:vii56–vii64, 2014.

[564] Craig, A. Personalised Medicine with Companion Diagnostics: The Intercept of Medicines and Medical Devices in the Regulatory Landscape - European Medical Journal. *EMJ Innov*, 1(1):47–53, January 2017.

[565] Vaske, C. J., Benz, S. C., Sanborn, J. Z., et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, 26(12):i237–i245, June 2010.

[566] Daemen, A., Griffith, O. L., Heiser, L. M., et al. Modeling precision treatment of breast cancer. *Genome biology*, 14(10):R110, October 2013.

[567] van der Velden, D. L., van Herpen, C. M. L., van Laarhoven, H. W. M., et al. Molecular Tumor Boards: current practice and future needs. *Biometrika*, September 2017.

[568] Bryce, A. H., Egan, J. B., Borad, M. J., et al. Experience with precision genomics and tumor board, indicates frequent target identification, but barriers to delivery. *Oncotarget*, 8(16):27145–27154, March 2017.

[569] Seeber, A., Gastl, G., Ensinger, C., et al. Treatment of patients with refractory metastatic cancer according to molecular profiling on tumor tissue in the clinical routine: an interim-analysis of the ONCO-T-PROFILE project. *Genes & Cancer*, October 2016.

[570] Parker, B. A., Schwaederle, M., Scur, M. D., et al. Breast Cancer Experience of the Molecular Tumor Board at the University of California, San Diego Moores Cancer Center. *Journal of Oncology Practice*, 11(6):442–449, November 2015.

[571] Oberg, J. A., Bender, J. L. G., Sulis, M. L., et al. Implementation of next generation sequencing into pediatric hematology-oncology practice: moving beyond actionable alterations. *Genome Medicine*, 8(1):1572, 2016.

[572] Lane, B. R., Bissonnette, J., Waldherr, T., et al. Development of a Center for Personalized Cancer Care at a Regional Cancer Center: Feasibility Trial of an Institutional Tumor Sequencing Advisory Board. *The Journal of Molecular Diagnostics*, 17(6):695–704, November 2015.

[573] Pincez, T., Clément, N., Lapouble, E., et al. Feasibility and clinical integration of molecular profiling for target identification in pediatric solid tumors. *Pediatric Blood & Cancer*, 64(6), November 2016.

[574] Foundation Medicine. https://www.foundationmedicine.com.

[575] CeGaT GmbH - Genetic Diagnostics and Sequencing Services. https://www.cegat.de/en/.

[576] Beltran, H., Eng, K., Mosquera, J. M., et al. Whole-Exome Sequencing of Metastatic Cancer and Biomarkers of Treatment Response. *JAMA Oncology*, 1(4):466–474, July 2015.

[577] Worst, B. C., van Tilburg, C. M., Balasubramanian, G. P., et al. Next-generation personalised medicine for high-risk paediatric cancer patients – The INFORM pilot study. *European Journal of Cancer*, 65:91–101, September 2016.

[578] Good, B. M., Ainscough, B. J., McMichael, J. F., et al. Organizing knowledge to enable personalization of medicine in cancer. *Genome biology*, 15(8):1113, August 2014.

[579] World Health Organization. Factsheet Cancer Germany 2018 https://gco.iarc.fr/today/data/factsheets/populations/276-germany-fact-sheets.pdf. Technical report, May 2019.

[580] BreastCancerOrg. U.S. Breast Cancer Statistics. https://www.breastcancer.org/symptoms/ understand_bc/statistics.

[581] Robert Koch Institut and Statistisches Bundesamt. Gesundheitsberichterstattung des Bundes: Brustkrebs. https://www.rki.de/EN/Content/Health_Monitoring/Health_Reporting/GBEDownloadsT/ brustkrebs.pdf . May 2005.

[582] Hon, J. D. C., Singh, B., Sahin, A., et al. Breast cancer molecular subtypes: from TNBC to QNBC. *American Journal of Cancer Research*, 6(9):1864, 2016.

[583] Sørlie, T., Perou, C. M., Tibshirani, R., et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, 98(19):10869–10874, September 2001.

[584] Rivenbark, A. G., O'Connor, S. M., and Coleman, W. B. Molecular and Cellular Heterogeneity in Breast Cancer: Challenges for Personalized Medicine. *The American Journal of Pathology*, 183(4):1113–1124, October 2013.

[585] Yersal, O. Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World Journal of Clinical Oncology*, 5(3):412–424, 2014.

[586] Dai, X., Li, T., Bai, Z., et al. Breast cancer intrinsic subtype classification, clinical use and future trends. *American Journal of Cancer Research*, 5(10):2929, 2015.

[587] American Cancer Society. Treatment of Breast Cancer by Stage. https://www.cancer.org/cancer/breast-cancer/treatment/treatment-of-breast-cancer-by-stage.html, 2018.

[588] Lillian Smyth, C. H. Adjuvant hormonal therapy in premenopausal women with breast cancer. *Indian Journal of Medical and Paediatric Oncology*, 36(4):195, 2015.

[589] Mokbel, R., Karat, I., and Mokbel, K. Adjuvant endocrine therapy for postmenopausal breast cancer in the era of aromatase inhibitors: an update. *International Seminars in Surgical Oncology*, 3(1):31, 2006.

[590] Jennifer L Hsu, M.-C. H. The role of HER2, EGFR, and other receptor tyrosine kinases in breast cancer. *Cancer metastasis reviews*, 35(4):575–588, December 2016.

[591] Lehmann, B. D., Bauer, J. A., Chen, X., et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of Clinical Investigation*, 121(7):2750–2767, July 2011.

[592] McCann, K. E., Hurvitz, S. A., and McAndrew, N. Advances in Targeted Therapies for Triple-Negative Breast Cancer. *Drugs*, 79(11):1217–1230, June 2019.

[593] Jhan, J.-R. and Andrechek, E. R. Effective personalized therapy for breast cancer based on predictions of cell signaling pathway activation from gene expression analysis. *Oncogene*, 36(25):3553–3561, June 2017.

[594] Ravdin, P. M., Siminoff, L. A., Davis, G. J., et al. Computer Program to Assist in Making Decisions About Adjuvant Therapy for Women With Early Breast Cancer. *Journal of Clinical Oncology*, 19(4):980–991, February 2001.

[595] dos Reis, F. J. C., Wishart, G. C., Dicks, E. M., et al. An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. *Breast cancer research*, 19(1):58, December 2017.

[596] Chen, L. L., Bush, D., Fong, A., et al. CancerMath.net: Web-Based Calculators for Breast Carcinoma. http://www.lifemath.net/cancer/breastcancer/therapy/index.php. 2008.

[597] Nik-Zainal, S., Davies, H., Staaf, J., et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, June 2016.

[598] Yates, L. R. and Desmedt, C. Translational Genomics: Practical Applications of the Genomic Revolution in Breast Cancer. *Clinical Cancer Research*, 23(11):2630–2639, June 2017.

[599] Parker, J. S., Mullins, M., Cheang, M. C. U., et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167, March 2009.

[600] Győrffy, B., Benke, Z., Lánczky, A., et al. RecurrenceOnline: an online analysis tool to determine breast cancer recurrence and hormone receptor status using microarray data. *Breast Cancer Research and Treatment*, 132(3):1025–1034, July 2011.

[601] Alcaraz, N., List, M., Batra, R., et al. De novo pathway-based biomarker identification. *Nucleic Acids Research*, 45(16), September 2017.

[602] Haibe-Kains, B., Desmedt, C., Loi, S., et al. A three-gene model to robustly identify breast cancer molecular subtypes. . *Journal of the National Cancer Institute*, 104(4):311–325, January 2012.

[603] Desmedt, C., Haibe-Kains, B., Wirapati, P., et al. Biological Processes Associated with Breast Cancer Clinical Outcome Depend on the Molecular Subtypes. *Clinical Cancer Research*, 14(16):5158–5165, August 2008.

[604] Wirapati, P., Sotiriou, C., Kunkel, S., et al. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast cancer research*, 10(4):R65, 2008.

[605] Alcaraz, N., List, M., Dissing-Hansen, M., et al. Robust de novo pathway enrichment with KeyPathwayMiner 5. *F1000Research*, 5:1531, 2016.

[606] Eo, H.-S., Heo, J. Y., Choi, Y., et al. A pathway-based classification of breast cancer integrating data on differentially expressed genes, copy number variations and MicroRNA target genes. *Molecules and Cells*, 34(4):393–398, September 2012.

[607] Mejía-Pedroza, R. A., Espinal-Enríquez, J., and Hernández-Lemus, E. Pathway-Based Drug Repositioning for Breast Cancer Molecular Subtypes. *Frontiers in Pharmacology*, 9:1040, August 2018.

[608] Ravdin, P. M., Siminoff, L. A., Davis, G. J., et al. Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. *Journal of Clinical Oncology*, 19(4):980–991, February 2001.

[609] Schneider, L., Stöckel, D., Kehl, T., et al. DrugTargetInspector: An assistance tool for patient treatment stratification. *International Journal of Cancer*, 138(7):1765–1776, April 2016.

[610] Schneider, L., Kehl, T., Thedinga, K., et al. ClinOmicsTrail-bc: a visual analytics tool for breast cancer treatment stratification. *Bioinformatics*, 35(24):5171–5181, April 2019.

[611] Slamon, D. J., Clark, G. M., Wong, S. G., et al. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*, 235(4785):177–182, January 1987.

[612] Forbes, S. A., Bhamra, G., Bamford, S., et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). *Current Protocols in Human Genetics*, 57(1):10.11.1–10.11.26, April 2008.

[613] Srihari, S., Kalimutho, M., Lal, S., et al. Understanding the functional impact of copy number alterations in breast cancer using a network modeling approach. *Molecular BioSystems*, 12(3):963–972, February 2016.

[614] Szyf, M., Pakneshan, P., and Rabbani, S. A. DNA methylation and breast cancer. *Biochemical Pharmacology*, 68(6):1187–1197, September 2004.

[615] Haber, D. A. and Settleman, J. Drivers and passengers. *Nature*, 446(7132):145–146, March 2007.

[616] Børresen-Dale, A.-L. TP53 and breast cancer. *Human Mutation*, 21(3):292–300, February 2003.

[617] Mukohara, T. PI3K mutations in breast cancer: prognostic and therapeutic implications. *Breast Cancer: Targets and Therapy*, May 2015.

[618] Usary, J., Llaca, V., Karaca, G., et al. Mutation of GATA3 in human breast tumors. *Oncogene*, 23(46):7669–7678, October 2004.

[619] Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., et al. IntOGen-mutations identifies cancer drivers across tumor types. *Nature Methods*, 10(11):1081–1082, September 2013.

[620] Wheeler, H. E., Maitland, M. L., Dolan, M. E., et al. Cancer pharmacogenomics: strategies and challenges. *Nature Reviews Genetics*, 14(1):23–34, January 2013.

[621] Goetz, M. P., Kamal, A., and Ames, M. M. Tamoxifen Pharmacogenomics: The Role of CYP2D6 as a Predictor of Drug Response. *Clinical pharmacology and therapeutics*, 83(1):160–166, January 2008.

[622] Giancotti, F. G. Deregulation of Cell Signaling in Cancer. *FEBS letters*, 588(16):2558–2570, August 2014.

[623] Velloso, F. J., Bianco, A. F., Farias, J. O., et al. The crossroads of breast cancer progression: insights into the modulation of major signaling pathways. *OncoTargets and therapy*, 10:5491–5524, 2017.

[624] Luo, M. and Guan, J.-L. Focal adhesion kinase: A prominent determinant in breast cancer initiation, progression and metastasis. *Cancer Letters*, 289(2):127–139, March 2010.

[625] Africander, D. and Storbeck, K.-H. Steroid metabolism in breast cancer: Where are we and what are we missing? *Molecular and Cellular Endocrinology*, 466:86–97, May 2018.

[626] Saura, C., Roda, D., Roselló, S., et al. A First-in-Human Phase I Study of the ATP-Competitive AKT Inhibitor Ipatasertib Demonstrates Robust and Safe Targeting of AKT in Patients with Solid Tumors. *Cancer Discovery*, 7(1):102–113, January 2017.

[627] Porta, C., Paglino, C., and Mosca, A. Targeting PI3K/Akt/mTOR Signaling in Cancer. *Frontiers in Oncology*, 4(Suppl 4):2905, 2014.

[628] Kutmon, M., Riutta, A., Nunes, N., et al. WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Research*, 44(D1):D488–D494, January 2016.

[629] Johnson, C. R. *Matrix Theory and Applications*, volume 40. American Mathematical Society, 1989.

[630] Grossman, R. L., Heath, A. P., Ferretti, V., et al. Toward a Shared Vision for Cancer Genomic Data. *International Journal of Cancer*, 375(12):1109–1112, September 2016.

[631] Wold, S., Esbensen, K., and Geladi, P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, August 1987.

[632] Maaten, L. v. d. and Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[633] Sanchez-Vega, F., Mina, M., Armenia, J., et al. Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell*, 173(2):321–337.e10, April 2018.

[634] Ramensky, V., Bork, P., and Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Research*, 30(17):3894, September 2002.

[635] Li, Y. H., Yu, C. Y., Li, X. X., et al. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Research*, 46(Database issue):D1121, January 2018.

[636] Twomey, J. D., Brahme, N. N., and Zhang, B. Drug-biomarker co-development in oncology - 20 years and counting. . *Drug Resistance Updates*, 30:48–62, January 2017.

[637] Dean, L. Trastuzumab (Herceptin) Therapy and ERBB2 (HER2) Genotype. *Drug Resistance Updates*, 30:48–62, August 2015.

[638] Zanger, U. M. and Schwab, M. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & Therapeutics*, 138(1):103–141, April 2013.

[639] Božina, N., Bradamante, V., and Lovrić, M. Genetic Polymorphism of Metabolic Enzymes P450 (CYP) as a Susceptibility Factor for Drug Response, Toxicity, and Cancer Risk. *Archives of Industrial Hygiene and Toxicology*, 60(2):269, July 2009.

[640] Choi, C.-H. ABC transporters as multidrug resistance mechanisms and the development of chemosensitizers for their reversal. *Cancer Cell International*, 5(1):30, 2005.

[641] Lepper, E. R., Nooter, K., Verweij, J., et al. Mechanisms of resistance to anticancer drugs: the role of the polymorphic ABC transporters ABCB1 and ABCG2. *Future Medicine*, 6(2):115–138, November 2005.

[642] Dai, C.-l., Tiwari, A. K., Wu, C.-P., et al. Lapatinib (Tykerb, GW572016) Reverses Multidrug Resistance in Cancer Cells by Inhibiting the Activity of ATP-Binding Cassette Subfamily B Member 1 and G Member 2. *Cancer Research*, 68(19):7905–7914, October 2008.

[643] Matsushita, H., Vesely, M. D., Koboldt, D. C., et al. Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting. *Nature*, 482(7385):400–404, February 2012.

[644] Fernando, J. and Kumar, S. Principles of cancer treatment by immunotherapy. *Surgery*, 33(3):117–121, March 2015.

[645] Lauss, M., Donia, M., Harbst, K., et al. Mutational and putative neoantigen load predict clinical benefit of adoptive T cell therapy in melanoma. *Nature Communications*, 8(1):1081, November 2017.

[646] MD Anderson Cancer Center. Human DNA Repair Genes. https://www.mdanderson.org/documents/Labs/Wood-Laboratory/human-dna-repair-genes.html, 2018.

[647] Schubert, B., de la Garza, L., Mohr, C., et al. ImmunoNodes – graphical development of complex immunoinformatics workflows. *BMC Bioinformatics*, 18(1):242, December 2017.

[648] Thiagarajan, P. S., Sinyuk, M., Turaga, S. M., et al. Cx26 drives self-renewal in triple-negative breast cancer via interaction with NANOG and focal adhesion kinase. *Nature Communications*, 9(1):578, February 2018.

[649] Golubovskaya, V. M., Ylagan, L., Miller, A., et al. High focal adhesion kinase expression in breast carcinoma is associated with lymphovascular invasion and triple-negative phenotype. *BMC cancer*, 14(1):769, December 2014.

[650] Quirke, V. M. Tamoxifen from Failed Contraceptive Pill to Best-Selling Breast Cancer Medicine: A Case-Study in Pharmaceutical Innovation. *Frontiers in Pharmacology*, 8:102, 2017.

[651] Early Breast Cancer Trialists Collaborative Group. Tamoxifen for early breast cancer: an overview of the randomised trials. *The Lancet*, 351(9114):1451–1467, May 1998.

[652] Schroth, W., Antoniadou, L., Fritz, P., et al. Breast Cancer Treatment Outcome With Adjuvant Tamoxifen Relative to Patient CYP2D6 and CYP3C19 Genotypes. *Journal of Clinical Oncology*, 25(33):5187–5193, November 2007.

[653] Ring, A. and Dowsett, M. Mechanisms of tamoxifen resistance. *Endocrine-Related Cancer*, 11(4):643–658, December 2004.

[654] Ascenzi, P., Bocedi, A., and Marino, M. Structure–function relationship of estrogen receptor α and β: Impact on human health. *Molecular Aspects of Medicine*, 27(4):299–402, August 2006.

[655] Periyasamy, M., Patel, H., Lai, C.-F., et al. APOBEC3B-Mediated Cytidine Deamination Is Required for Estrogen Receptor Action in Breast Cancer. *Cell Reports*, 13(1):108–121, October 2015.

[656] Martin, L.-A., Farmer, I., Johnston, S. R. D., et al. Elevated ERK1/ERK2/estrogen receptor cross-talk

enhances estrogen-mediated signaling during long-term estrogen deprivation. *Endocrine-Related Cancer*, 12(Supplement 1):S75–S84, July 2005.

[657] Fabian, C. J. The what, why and how of aromatase inhibitors: hormonal agents for treatment and prevention of breast cancer. *International Journal of Clinical Practice*, 61(12):2051–2063, December 2007.

[658] Francis, P. A., Regan, M. M., Fleming, G. F., et al. Adjuvant Ovarian Suppression in Premenopausal Breast Cancer. *New England Journal of Medicine*, 372(5):436–446, January 2015.

[659] Couzin-Frankel, J. Cancer Immunotherapy. *Science*, 342(6165):1432–1433, December 2013.

[660] Achkar, T. and Tarhini, A. A. The use of immunotherapy in the treatment of melanoma. *Journal of Hematology & Oncology*, 10(1):88, December 2017.

[661] Brahmer, J. R. and Pardoll, D. M. Immune Checkpoint Inhibitors: Making Immunotherapy a Reality for the Treatment of Lung Cancer. *Cancer Immunology Research*, 1(2):85–91, August 2013.

[662] Grillo-Lopez, A. J., White, C. A., Dallaire, B. K., et al. Rituximab The First Monoclonal Antibody Approved for the Treatment of Lymphoma. *Current Pharmaceutical Biotechnology*, 1(1):1–9, June 2000.

[663] Garassino, M. C., Cho, B.-C., Kim, J.-H., et al. Durvalumab as third-line or later treatment for advanced non-small-cell lung cancer (ATLANTIC): an open-label, single-arm, phase 2 study. *The Lancet. Oncology*, 19(4):521–536, April 2018.

[664] Motzer, R. J., Escudier, B., McDermott, D. F., et al. Nivolumab versus Everolimus in Advanced Renal-Cell Carcinoma. *New England Journal of Medicine*, 373(19):1803–1813, November 2015.

[665] Mouw, K. W., Goldberg, M. S., Konstantinopoulos, P. A., et al. DNA Damage and Repair Biomarkers of Immunotherapy Response. *Cancer Discovery*, 7(7):675–693, July 2017.

[666] Levine, A. J., Momand, J., and Finlay, C. A. The p53 tumour suppressor gene. *Nature*, 351(6326):453–456, June 1991.

[667] Gonzalo, S., García-Cao, M., Fraga, M. F., et al. Role of the RB1 family in stabilizing histone methylation at constitutive heterochromatin. *Nature Cell Biology*, 7(4):420–428, April 2005.

[668] Palmieri, G., Colombino, M., Cossu, A., et al. Genetic instability and increased mutational load: which diagnostic tool best direct patients with cancer to immunotherapy? *Journal of Translational Medicine*, 15(1):2189, 2017.

[669] Matsuoka, S., Ballif, B. A., Smogorzewska, A., et al. ATM and ATR Substrate Analysis Reveals Extensive Protein Networks Responsive to DNA Damage. *Science*, 316(5828):1160–1166, May 2007.

[670] Brown, J. S., Sundar, R., and Lopez, J. Combining DNA damaging therapeutics with immunotherapy: more haste, less speed. *British Journal of Cancer*, 118(3):312–324, February 2018.

[671] Schmidt, C. The benefits of immunotherapy combinations. *Nature*, 552(7685):S67–S69, December 2017.

[672] Wein, L., Luen, S. J., Savas, P., et al. Checkpoint blockade in the treatment of breast cancer: current status and future directions. *British Journal of Cancer*, 500:1, May 2018.

[673] Meagher, M. and Lightowlers, R. N. The role of TDP1 and APTX in mitochondrial DNA repair. *Biochimie*, 100:121–124, May 2014.

[674] Santos, L. S., Gomes, B. C., Gouveia, R., et al. The role of CCNH Val270Ala (rs2230641) and other nucleotide excision repair polymorphisms in individual susceptibility to well-differentiated thyroid cancer. *Oncology Reports*, 30(5):2458–2466, November 2013.

[675] Kitao, H., Yamamoto, K., Matsushita, N., et al. Functional Interplay between BRCA2/FancD1 and FancC in DNA Repair. *The Journal of Biological Chemistry*, 281(30):21312–21320, July 2006.

[676] Kowal, P., Gurtan, A. M., Stuckert, P., et al. Structural Determinants of Human FANCF Protein That Function in the Assembly of a DNA Damage Signaling Complex. *The Journal of Biological Chemistry*, 282(3):2047–2055, January 2007.

[677] Kolodner, R. D. and Marsischky, G. T. Eukaryotic DNA mismatch repair. *Current Opinion in Genetics & Development*, 9(1):89–96, February 1999.

[678] Wilson, D. M. *Base Excision Repair Pathway, The: Molecular Mechanisms And Role In Disease Development And Therapeutic Design*. World Scientific, January 2017.

[679] Yasuhara, T., Suzuki, T., Katsura, M., et al. Rad54B serves as a scaffold in the DNA damage response that limits checkpoint strength. *Nature Communications*, 5(1):5426, November 2014.

[680] Nakatsu, Y., Asahina, H., Citterio, E., et al. XAB2, a Novel Tetratricopeptide Repeat Protein Involved in Transcription-coupled DNA Repair and Transcription. *The Journal of Biological Chemistry*, 275(45):34931–34937, November 2000.

[681] Saiki, R. K., Walsh, P. S., Levenson, C. H., et al. Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes. *PNAS*, 86:6230–6234, August 1989.

[682] Erlich, H. A., Opelz, G., and Hansen, J. HLA DNA Typing and Transplantation. *Immunity*, 14(4):347–356, April 2001.

[683] Szolek, A., Schubert, B., Mohr, C., et al. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics*, 30(23):3310–3316, December 2014.

[684] Andreatta, M. and Nielsen, M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics*, 32(4):511–517, February 2016.

[685] Magner, L. N. and Kim, O. J. *A History of Medicine*. CRC Press, December 2017.

[686] Prakash, V., Moore, M., and Yáñez-Muñoz, R. J. Current Progress in Therapeutic Gene Editing for Monogenic Diseases. *Molecular Therapy*, 24(3):465–474, March 2016.

[687] Goossens, N., Nakagawa, S., Sun, X., et al. Cancer biomarker discovery and validation. *Translational cancer research*, 4(3):256, June 2015.

[688] Hudis, C. A. Trastuzumab — Mechanism of Action and Use in Clinical Practice. *New England Journal of Medicine*, 357(1):39–51, 2007.

[689] Strom, S. P. Current practices and guidelines for clinical next-generation sequencing oncology testing. *Cancer Biology & Medicine*, 13(1):3, March 2016.

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

[691] Gomez-Cabrero, D., Abugessaisa, I., Maier, D., et al. Data integration in the era of omics: current and future challenges. *BMC systems biology*, 8(Suppl 2):I1, 2014.

[692] Zanzoni, A., Soler-López, M., and Aloy, P. A network medicine approach to human disease. *FEBS letters*, 583(11):1759–1765, June 2009.

[693] Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., et al. Delineation of prognostic biomarkers in prostate cancer. *Nature*, 412(6849):822–826, August 2001.

[694] Singer, J., Irmisch, A., Ruscheweyh, H.-J., et al. Bioinformatics for precision oncology. *Briefings in Bioinformatics*, May 2019.

[695] Marian, A. J. Challenges in Medical Applications of Whole Exome/Genome Sequencing Discoveries. *Trends in Cardiovascular Medicine*, 22(8):219–223, November 2012.

[696] EMBL-EBI. Ensembl Variation. https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html. Technical report, January 2019.

[697] Love, M. I., Huber, W., and Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):1–21, December 2014.

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

[699] Risso, D., Ngai, J., Speed, T. P., et al. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, 32(9):896–902, August 2014.

[700] Efron, B. and Tibshirani, R. On Testing the Significance of Sets of Genes. *The Annals of Applied Statistics*, 1(1):107–129, June 2007.

[701] Ackermann, M. and Strimmer, K. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10(1):1–20, February 2009.

[702] Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, December 1945.

[703] NIA. National Institute on Aging. http://www.nia.nih.gov/. Technical report, 2015.

[704] Nishimura, D. A View From the Web: Biocarta. *Biotech Software Internet Report*, 2(3):117–120, July 2004.

[705] NCI. National Cancer Institute. http://www.cancer.gov/. Technical report, 2015.

[706] Frolkis, A., Knox, C., Lim, E., et al. SMPDB: The Small Molecule Pathway Database. *Nucleic Acids Research*, 38(Database issue):D480–487, January 2010.

[707] Meyer, L. R., Zweig, A. S., Hinrichs, A. S., et al. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Research*, 41(D1):D64–D69, January 2013.

[708] Karolchik, D., Barber, G. P., Casper, J., et al. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Research*, 42(D1):D764–D770, January 2014.

[709] Xiao, F., Zuo, Z., Cai, G., et al. miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Research*, 37(Database issue):D105–110, January 2009.

[710] Li, J.-H., Liu, S., Zhou, H., et al. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Research*, 42(Database issue):D92–97, January 2014.

[711] Krek, A., Grün, D., Poy, M. N., et al. Combinatorial microRNA target predictions. *Nature Genetics*, 37(5):495–500, May 2005.

[712] Garcia, D. M., Baek, D., Shin, C., et al. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nature Structural & Molecular Biology*, 18(10):1139–1146, October 2011.

[713] Kamburov, A., Stelzl, U., Lehrach, H., et al. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Research*, 41(D1):D793–D800, November 2012.

[714] Finn, R. D., Bateman, A., Clements, J., et al. Pfam: the protein families database. *Nucleic Acids Research*, 42(Database issue):D222–230, January 2014.

[715] Benjamini, Y. and Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, August 2001.

[716] Bui, H.-H., Sidney, J., Peters, B., et al. Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics*, 57(5):304–314, May 2005.

[717] Parker, K. C., Bednarek, M. A., and Coligan, J. E. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *Journal of Immunology*, 152(1):163–175, January 1994.

[718] Sidney, J., Assarsson, E., Moore, C., et al. Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries. *Immunome Research*, 4(1):2, 2008.

[719] Nielsen, M., Lundegaard, C., and Lund, O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics*, 8(1):238, December 2007.

[720] Karosiene, E., Rasmussen, M., Blicher, T., et al. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics*, 65(10):711–724, July 2013.

[721] Zhang, H., Lund, O., and Nielsen, M. The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics*, 25(10):1293–1299, May 2009.

[722] Peters, B. and Sette, A. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics*, 6(1):132, December 2005.

[723] Kim, Y., Sidney, J., Pinilla, C., et al. Derivation of an amino acid similarity matrix for peptide:MHC binding and its application as a Bayesian prior. *BMC Bioinformatics*, 10(1):394, December 2009.

[724] Dönnes, P. and Elofsson, A. Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics*, 3(1):25, December 2002.

[725] Rammensee, H. G., Bachmann, J., Emmerich, N. P. N., et al. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, 50(3-4):213–219, November 1999.

[726] Toussaint, N. C., Feldhahn, M., Ziehm, M., et al. T-cell epitope prediction based on self-tolerance. *ACM*, pages 584–588, August 2011.

[727] Parker, J. S., Mullins, M., Cheang, M. C. U., et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167, March 2009.

A

Supplementary Material

This Appendix contains additional information and details on various aspects of the methods, tools, and analyses presented in the main part of this thesis.

A.1 File formats

The tools and methods presented in the course of this thesis require different types of standardized file formats as input, some of the more complex ones are presented in the following sections.

A.1.1 FASTA

The FASTA file format is used to describe nucleotide or peptide sequences. A sequence in FASTA format begins with a single-line description (starting with '>'), followed by several lines of sequence data. The bases (or amino acids) are represented by nucleotides from the alphabet {A,C, G,T} (or single-letter amino acid codes). An example sequence in FASTA format could look like this:

```
>Sequence_A
GGTAAGTCCTCTAGTACAAACACCCCCAATATTGTGATATAATTAAAATTATATTCATAT
TCTGTTGCCAGAAAAAACACTTTTAGGCTATATTAGAGCCATCTTCTTTGAAGCGTTGTC
>Sequence_B
GGTAAGTGCTCTAGTACAAACACCCCCAATATTGTGATATAATTAAAATTATATTCATAT
TCTGTTGCCAGATTTTACACTTTTAGGCTATATTAGAGCCATCTTCTTTGAAGCGTTGTC
TATGCATCGATCGACGACTG
```

A.1.2 FASTQ

The FASTQ file format is an extension of the FASTA file format, which contains additional quality information for every base in a nucleotide sequence.

Each entry in a FASTQ file is composed of four lines, where each line contains different types of information. The first line starts with a sequence identifier (starting with '@'), followed by optional descriptions. The second line contains the actual sequence. The third line oftentimes only contains a '+' and visually separates the sequence from the fourth line, which contains

Phred quality scores (cf. **Section 3.1.2.3**) for every nucleotide using ASCII characters. An example of a file in FASTQ format could look like this:

```
@Sequence_A
GGTAAGTCCTCTAGTACAAACACCCCCAATATTGTGATATAATTAAAATTATATTCATAT...
+
ABBBBFFFFFFFGGGGGGGGGGHHHHHHHHHHHGHHFGHHFHGHHHHHGGEHFCGGHDGFB...
@Sequence_B
GGTAAGTGCTCTAGTACAAACACCCCCAATATTGTGATATAATTAAAATTATATTCATAT....
+
AABBBFFFFFFBGGGGGGGGGGGGFEEGGHHHHHHHFHHFEGHFBEGFHHHHGHHHHHHH...
```

A.1.3 SAM / BAM

The Sequence Alignment/Map (SAM) format is a text-based format to store sequence alignments, typically between a sequence of interest and a predefined reference. SAM files consist of a header section and an alignment section. Header lines are marked by '@' and typically contain metadata on the performed analyses. Each aligned fragment is represented by a row with at least eleven columns, where the first eleven columns contain the information listed in **Table A.1**.

Column	Field name	Description
1	QNAME	Read name
2	FLAG	Record's flag
3	RNAME	Reference name
4	POS	1-based position on reference
5	MAPQ	Mapping quality
6	CIGAR	Compact Idiosyncratic Gapped Alignment Report of alignment
7	RNEXT	Reference of the next mate/segment
8	PNEXT	Position of the next mate/segment
9	TLEN	Observed length of template
10	SEQ	Read sequence
11	QUAL	ASCII-encoded Phred base qualities

Table A.1 **Description of fields in SAM format**. Please refer to the following website for the full format specification: https://samtools.github.io/hts-specs/SAMv1.pdf.

An example file in SAM format could look like this:

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
@SQ SN:ref2 LN:40
r001 163 ref 7 30 8M4I4M1D3M = 37 39 TTAGATAAAGAGGATACTG *
r002 0 ref 9 30 1S2I6M1P1I1P1I4M2I * 0 0 AAAAGATAAGGGATAAA *
r003 0 ref 9 30 5H6M * 0 0 AGCTAA *
r004 0 ref 16 30 6M14N1I5M * 0 0 ATAGCTCTCAGC *
```

The binary equivalent to the SAM format is the Binary Alignment/Map (BAM) format.

A.1.4 VCF

The Variant Call Format (VCF) is the preferred format to represent variation data. VCF files are tab-delimited text files where each variation is given in one row of eight predefined columns (cf. **Table A.2**), followed by the mutation status in one or several samples. Metadata on the performed analyses can be specified in header lines (starting with '##').

Column	Field name	Description
1	#CHROM	Chromosome
2	POS	Coordinate - the start of the variant
3	ID	Identifier
4	REF	Reference allele (in the reference genome)
5	ALT	Alternative allele (allele found in the sample under investigation)
6	QUAL	Score - a quality score
7	FILTER	PASS / FAIL - if the variant passed quality filters
8	INFO	Further information - keys in the INFO fields can be defined in header lines above the table
9	FORMAT	Information about the following columns (e.g., containing GT/DP, HET, HOM, RD, AD, AF)

Table A.2 **Description of fields in VCF format**. Please refer to the following website for the full format specification: https://samtools.github.io/hts-specs/VCFv4.2.pdf.

An example file in VCF format could look like this:

```
##fileformat=VCFv4.0
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT Sample1 Sample2
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ
0|0:48:1:51,51 1|0:48:8:51,51
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50
0|1:3:5:65,3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB
GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2
```

A.1.5 BED

The BED file format is used to describe genomic regions. In this format, a tab-delimited text file is used that contains one feature per line. Each line contains at least three values: the name of the chromosome (either just the number or with an additional 'chr' prefix), the start position of the feature in standard chromosomal coordinates, and the end position of the feature in standard chromosomal coordinates. The BED format is, for example, used to describe regions of open chromatin (cf. **Section 4.3.2**). When used to display genomic regions in a genome browser, additional columns can be included. These columns contain additional layout information like the feature's display name and drawing options. An example file in BED format could look like this:

```
chr1 213941196 213942363
chr1 213942363 213943530
chr1 213943530 213944697
chr2 158364697 158365864
chr3 127477031 127478198
```

A.1.6 SEG

The SEGmented data (SEG) file format is a tab-delimited text-based format that is used to describe genomic locations. The first line contains a header describing the composition of each entry. Besides an optional identifier, the chromosome number and the start and end of a region of interest are given in terms of their genomic locations. When SEG files are used to store copy number alterations, these are commonly provided as another column containing log-ratios of the sample's copy number in this genomic region in comparison to the copy number in a reference sample or group. An example file in SEG format could look like this:

```
ID chrom loc.start loc.end seg.mean
GenomeWideSNP_416532 1 51598 76187 -0.7116
GenomeWideSNP_416532 1 76204 16022502 -0.029
GenomeWideSNP_416532 1 16026084 16026512 -2.0424
GenomeWideSNP_416532 1 16026788 17063449 -0.1024
GenomeWideSNP_416532 1 17067742 17134834 -0.6868
```

A.2 Variant Effect Predictor

Ensembl's Variant Effect Predictor (VEP) [220] is a variant annotation tool that categorizes variants based on their predicted impact on protein function. VEP distinguishes a large variety of so-called 'consequence types', i.e. different effects a mutation can have on the genomic location or protein it occurs in. **Figure A.1** gives an overview of the major consequence types considered in our analyses.

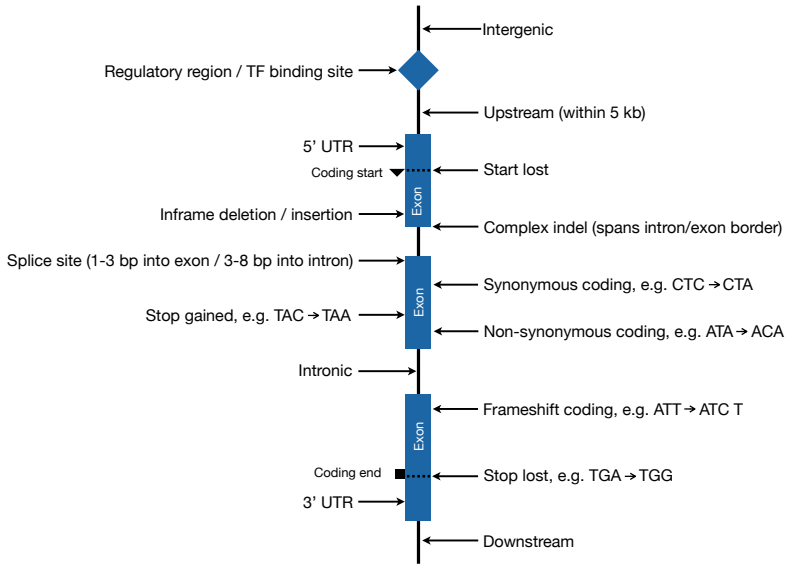


Figure A.1 Consequence terms in Variant Effect Predictor. Schematic overview on variant consequences as annotated by Variant Effect Predictor with respect to their genomic location. Figure adapted from [696]. Please refer to the following website for the complete specification: https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html.

A.3 Supplements for Graviton

Graviton is a general framework for the implementation of web-based, integrative, multi-omics systems-biology tools, which serves as the basis for our specialized analysis pipelines (cf. **Section 4.1**).

Identifier type	Reference
RefSeq	[268]
NCBI EntrezGene	[270]
HGNC symbols and IDs	[272]
Ensembl	[273]
UniProt	[274]
KEGG	[282]
miRBase	[278]

Table A.3 Excerpt of identifier types supported by Graviton. Identifier types for *Homo sapiens* supported by Graviton and its derived tools. Please refer to https://genetrail2.bioinf.uni-sb.de/mappings.html for a complete list of supported identifier types across species.

Statistic	Reference
(Log) mean fold quotient	[391]
Standard score (z-score)	[392]
Independent shrinkage t-test	[393]
Independent Student's t-test	[394]
Welch's t-test	[317]
Wilcoxon- Mann-Whitney test	[395]
Signal-to-noise ratio	[396]
F-test	[397]
Pearson correlation	[343]
Spearman correlation	[481]
DESeq2	[697]
EdgeR	[698]
RUVSeq	[699]

Table A.4 Entity-level statistics provided by Graviton. List of scoring methods for computation of scores of differential expression and methylation.

A.3.1 RESTful API usage

Using the RESTful API offered by Graviton, analyses can also be programmatically performed. A code snipped as an example of running a Gene Set Enrichment Analysis using GeneTrail2 in R is provided in **Listing A.1**.

A.4 Supplements for GeneTrail2

GeneTrail2 is a web service for the integrated analysis of genomics, transcriptomics, miRNomics, and proteomics data sets (cf. **Section 4.2**). Overviews of the provided enrichment methods, predefined biological categories, and alternative result visualizations are provided in **Sections A.4.1** to **A.4.3**.

A.4.1 Enrichment algorithms

The enrichment algorithms provided by GeneTrail2 consist of methods for the computation of the enrichment itself (cf. **Table A.5**) and methods for adjusting p-values in the case of multiple hypothesis testing (cf. **Table A.6**).

Enrichment method	Reference
Over-Representation Analysis	[325]
Weighted Gene Set Enrichment Analysis	[326]
Gene Set Enrichment Analysis	[327]
Two Sample *t*-Test	[393]
One-Sample *t*-Test	[397]
Max-Mean Statistic	[700]
Mean of Single Gene Statistic	[701]
Median of Single Gene Statistic	[701]
Sum of Single Gene Statistic	[701]
Wilcoxon Rank-Sum Test	[702]

Table A.5 **Enrichment algorithms provided by GeneTrail2**. This table lists the enrichment algorithms provided by the web service and their corresponding publications.

```
#request session
session.response <- GET("http://localhost:8080/Graviton/api/session")
session.id <- content(session.response)$session

#upload gene expression file
gene.expression.file <- "<path to file>"
upload.url <- paste0("http://genetrail2.bioinf.uni-sb.de/api/upload/matrix?
                    session=", session.id)

upload.response <- POST(upload.url,
       body=list(file=upload_file(gene.expression.file)), encode='multipart')
gene.expression.resourceId <- content(upload.response)$results$result$id

#define sample and reference set
sg <- toJSON("<list of sample names>")
rg <- toJSON("<list of reference names>")

#setup scoring
scoring.url <- paste0("http://genetrail2.bioinf.uni-sb.de/api/job/setup/scoring?
         session=", session.id)
job.setup.response <- POST(scoring.url,
                     body=list(method='independent-shrinkage-t-test',
                     sg=sg, rg=rg, file1=gene.expression.resourceId),
                     encode='form')

#run scoring
job.start.response <- GET(paste0("http://genetrail2.bioinf.uni-sb.de/api/job/start?
                   session=", session.id))

# query result
job.query.response <- GET(paste0("http://genetrail2.bioinf.uni-sb.de/api/job/query?
                     session=", session.id))
scores.resource.id <- content(job.query.response)$results$scores$id

#setup gsea
gsea.url <- paste0("http://genetrail2.bioinf.uni-sb.de/api/job/setup/gsea?
    session=", session.id)
gsea.setup.response <- POST(gsea.url,
                       body=list(significance=0.05, adjustment="benjamini_yekutieli",
                       categories= "[\"9606-gene-kegg-pathways\"]",
                       minimum=2, maximum=700, adjustSeparately=T,
                       input=scores.resource.id), encode='form')

gsea.start.response <- GET(paste0("http://genetrail2.bioinf.uni-sb.de/api/job/start?
                      session=", session.id))
gsea.query.response <- GET(paste0("http://genetrail2.bioinf.uni-sb.de/api/job/query?
                      session=", session.id))
gsea.enrichment.id <- content(gsea.query.response)$results$enrichment$id

#access the results
enrichment.results <- GET(paste0("http://genetrail2.bioinf.uni-sb.de/api/resource/",
                     "enrichment/", gsea.enrichment.id, "?session=", session.id,
                     "&categoryName=", "KEGG_-_Pathways",
                     "&significance=", "0.05"))
```

Listing A.1 **Using Graviton RESTful API.** Code example of how to run a GeneTrail2 enrichment analysis in R.

P-value adjustment method	Reference
Bonferroni	[306]
Šidák	[307]
Holm	[308]
Finner	[309]
Benjamini-Hochberg	[310]
Benjamini-Yekutieli	[311]

Table A.6 **P-value adjustment methods provided by GeneTrail2**. This table lists the p-value adjustment methods provided by the web service and their corresponding publications.

A.4.2 Predefined biological categories

GeneTrail2 provides more than 46,000 biological categories collected from over 30 databases (cf. **Table A.7**). Moreover, custom user categories can be uploaded to GeneTrail2 in Gene Matrix Transposed (`GMT`) file format (cf. **Section A.4.2.1**).

Type of category	Provided databases
Ontology and phenotype	Gene Ontology [424], National Institut on Aging DB [703]
Pathways	BioCarta [704], KEGG [282], National Cancer Institut DB [705], PharmGKB [518], Reactome [425], Small Molecule Pathway Database [706], WikiPathways [426]
Genomic positions	HG19 GRCh37 [707], HG19 GRCh38 [708]
Targets	DrugBank [427], TRANSFAC [539], mirDB [428], mi-Records [709], miRTarBase [538], StarBase [710], PicTar [711], TargetScan [712]
Collections and others	ConsensusPathDB [713], Protein families DB (Pfam) [714]

Table A.7 **Biological categories predefined in GeneTrail2**. This table lists the provided types of categories and their respective databases, including references to the corresponding publications. Please refer to https://genetrail2.bioinf.uni-sb.de/categories.html for a complete list of supported databases across species.

A.4.2.1 GMT file format

In the Gene Matrix Transposed (`GMT`) file format, every line represents a category (i.e., a gene set). Each line is divided into columns by tab characters. The first column corresponds to the name of the category and the second column to an optional description or URL. Each subsequent column defines a category member [415]:

```
CategoryA http://test.url/A GeneA GeneB GeneC GeneD
CategoryB http://test.url/B GeneA GeneD
CategoryC http://test.url/C GeneD GeneE GeneH
```

A.4.3 Additional enrichment views

GeneTrail2 provides various views for the enrichment results. Besides the default view (cf. **Figure 4.4**) of a list of enriched or depleted pathways, GeneTrail2 also provides an *inverse enrichment* view (cf. **Figure A.2**). Here, differentially expressed genes are listed in decreasing order of their score of deregulation. For each gene, the pathways and gene sets the gene belongs to are listed and they can be investigated with respect to their enrichment status. For the integrative analysis of enrichment results from multiple omics data sets, GeneTrail2's *comparative enrichment* view can be used (cf. **Figure A.3**). This specialized view allows comparing several enrichment results side-by-side. Currently, there are two modes for comparison: intersection and union. The intersection mode only displays categories that are significantly enriched in all performed enrichment analyses. The union mode displays any category that is significantly enriched at least once. For the visual analytics-based investigation of dependencies between enriched or depleted categories, GeneTrail2 also offers a *dependency wheel* visualization, which provides a circular representation of altered categories with connecting ribbons indicating the number of shared genes between two categories (cf. **Figure A.4**).

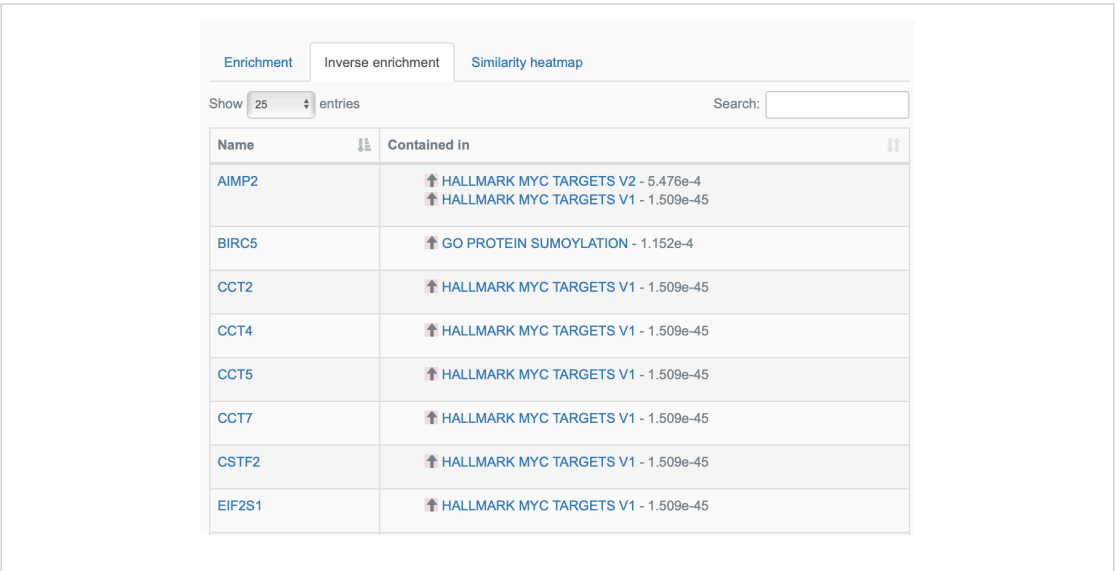


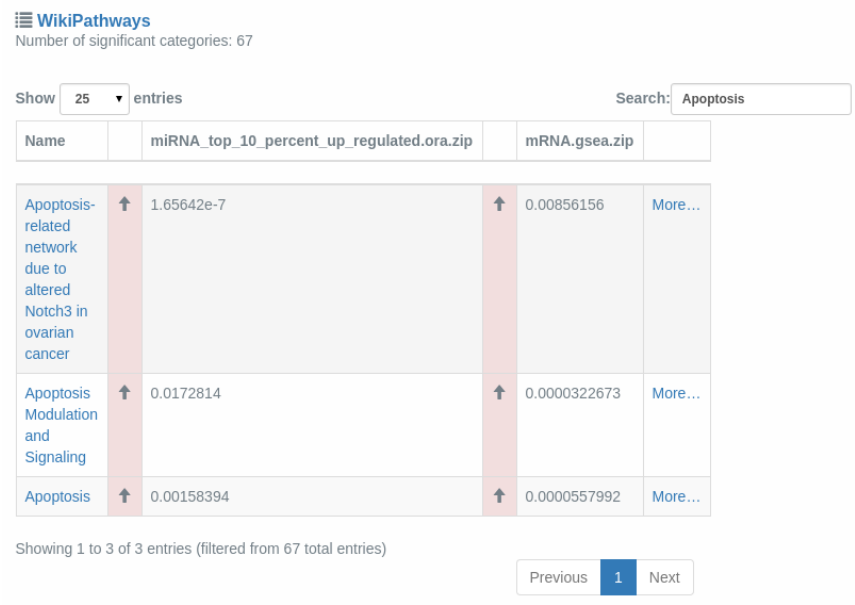
Figure A.2 **Inverse enrichment view**. For each gene in the considered set, all significantly enriched (or depleted) categories are listed with their corresponding adjusted p-values. Results for the data set discussed in **Section 4.2.3**.

A.5 Supplements for RegulatorTrail

RegulatorTrail is a web service for the identification of aberrant transcriptional regulators that are involved in pathogenic processes (cf. **Section 4.3**). Additional information for the two case studies we used RegulatorTrail for is provided in **Sections A.5.1** and **A.5.2**.

A.5.1 Breast cancer case study

In order to compare the capabilities of REGGAE with competing methods, we applied REGGAE and seven other methods (CSA, RIF1, RIF2, TDD, TED, TFactS, and TFRank, see **Section 4.3.1**) to a breast cancer data set and investigated whether we could identify key regulatory factors involved in breast cancer initiation and progression.



WikiPathways
Number of significant categories: 67

Show 25 entries Search: Apoptosis

Name	miRNA_top_10_percent_up_regulated.ora.zip	mRNA.gsea.zip	
Apoptosis-related network due to altered Notch3 in ovarian cancer	1.65642e-7	0.00856156	More…
Apoptosis Modulation and Signaling	0.0172814	0.0000322673	More…
Apoptosis	0.00158394	0.0000557992	More…

Showing 1 to 3 of 3 entries (filtered from 67 total entries)

Previous 1 Next

Figure A.3 Comparative enrichment view. Exemplary result of a comparative enrichment analysis (intersection mode). Categories enriched for both analyzed data sets (`miRNA_top_10_percent_up_regulated.ora.zip` and `mRNA.gsea.zip`) are listed with their respective adjusted p-values. Details on the analysed data sets can be found here: `https://genetrail2.bioinf.uni-sb.de/help?topic=integrative_analysis_wilms_mrna_mirna`.

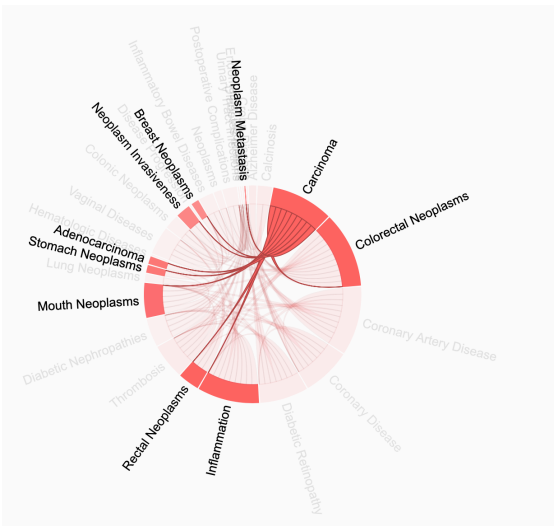


Figure A.4 Dependency wheel visualization. Exemplary result of a *dependency wheel* visualization for data set discussed in **Section 4.2.3** and enrichment of NIA phenotypes [703]. The width of connecting ribbons indicates the number of shared genes between two categories.

A.5.1.1 Sample groups

In this section, we describe the breast cancer data set published by Heiser *et al.* [489]. The data set contains gene expression profiles of 46 breast cancer cell lines. We obtained the status of the estrogen receptor (ER) for each cell line from a study by Neve *et al.* [490]. After the assignment, we obtained five distinct sample groups (cf. **Table A.8**).

Group	Description	Samples
1	Estrogen-receptor positive (ER+)	600MPE, BT474, BT483, CAMA1, HCC1428, LY2, MCF7, MDAMB134VI, MDAMB175VII, MDAMB361, MDAMB415, T47D, UACC812, ZR751, ZR7530, ZR75B
2	Estrogen-receptor negative (ER-)	AU565, BT20, BT549, HCC38, HCC70, HCC202, HCC1143, HCC1187, HCC1937, HCC1954, HCC2185, HCC3153, HS578T, MCF10A, MCF12A, MDAMB157, MDAMB231, MDAMB453, SKBR3, SUM225CWN, SUM1315MO2
3	Presumably estrogen-receptor positive (ER[+])	SUM52PE
4	Presumably estrogen-receptor negative (ER[-])	SUM149PT, SUM159PT
5	No information available (NA)	184B5, HCC1395, HCC1419, HCC1806, MCF10F, SUM185PE

Table A.8 **Sample groups for REGGAE breast cancer case study**. In all analyses presented in **Section 4.3.3**, we compared ER+ (Group 1) to ER- cells lines (Group 2).

A.5.1.2 Parameters and results

Correlation set analysis (CSA): For the Correlation Set Analysis [460], we used the implementation provided by the RegulatorTrail web service (cf. **Section 4.3**). For the five lists containing upregulated genes, we calculated an upper-tailed p-value. All p-values were estimated using a permutation test with 1,000,000 random permutations and an additional pseudo-count.

REGGAE: For the REGGAE analysis, we used the implementation provided by the RegulatorTrail web service (cf. **Section 4.3**). For each of the five lists, we sorted the lists decreasingly (with respect to their *t*-scores). As described in **Section 4.3.3**, we first computed Pearson's correlation coefficients between all genes and all associated regulators, and based on this information, we built the associated regulator lists (sorted decreasingly with respect to their association scores). The resulting p-values are adjusted using the Benjamini and Yekutieli method [311]. Finally, we performed an enrichment analysis using the Wilcoxon rank-sum test to detect the most influential regulators. All REGGAE analyses were performed using 1,000 random bootstrap replications.

RIF1 and RIF2: For the RIF1 and RIF2 analysis [459], we used the implementation provided by the RegulatorTrail web service (cf. **Section 4.3**). We used Pearson's correlation coefficients to compute the differences in correlation (between the two groups of interest) for each regulator and its target genes in the analyzed gene lists and the fold changes to assess differential expression.

TDD: For the TDD analysis [457], we implemented a Python script that calculates the respective statistic.

TED: In order to perform the analysis proposed by Yang *et al.* [457], we used the binomial test implemented in the RegulatorTrail web service (cf. **Section 4.3**). As a reference set, we used all genes that are targeted by a regulator in the used collection of RTIs. The resulting p-values are adjusted using the Benjamini and Yekutieli method [311].

TFactS: In order to perform the analysis proposed by Essaghir *et al.* [456], we used the hypergeometric test implemented in the RegulatorTrail web service (cf. **Section 4.3**). As a reference set, we used all genes that are targeted by a regulator in the used collection of RTIs. The resulting p-values are adjusted using the Benjamini and Yekutieli method [715].

TFRank: For the TFRank analysis [461], we used the prototype implementation provided on the authors' website (`http://web.tecnico.ulisboa.pt/aplf/code/tfrank/`). We used the unweighted network given by our collection of RTIs and the standard parameters also provided on the authors' website.

For each method, the complete results for the five gene sets described in **Section 4.3.3**, as well as their respective aggregated results can be accessed via the following link:

Click here to access / download the supplementary file from
www.lara-schneider.de/dissertation/REGGAE_BRCA_Case_study_full_results.xlsx

Method	Runtime [s]
CSA*	450.27 ($\pm$ 78.76)
REGGAE**	174.98 ($\pm$ 1.69)
REGGAE (without bootstrapping)	23.40 ($\pm$ 0.36)
RIF1	23.60 ($\pm$ 0.28)
RIF2	23.85 ($\pm$ 0.10)
TDD	14.86 ($\pm$ 0.63)
TED	658.20 ($\pm$ 29.80)
TFactS	42.37 ($\pm$ 0.23)
TFRank	116.74 ($\pm$ 4.22)

Table A.9 Runtime comparison for top 250 upregulated genes in REGGAE breast cancer case study. Note: Runtimes were obtained on an Intel Core i7-3770 processor. *CSA analysis was conducted using 1,000,000 permutations. **REGGAE analysis was performed using 1,000 bootstrap runs. Please note that a major part of the computation time of REGGAE (without bootstrapping) is spent on reading-in the large database of regulator-target interactions.

A.5.2 Wilms tumor case study

In a second case study, we used REGGAE to identify transcriptional regulators that potentially explain the differences between blastemal and non-blastemal Wilms tumors (cf. **Section 4.3.4.2**). To this end, 33 Wilms tumor samples were analyzed (cf. **Table A.12**). The following section contains the list of the top 50 regulators identified by REGGAE (cf. **Table A.10**) and a link to download all identified regulators.

Upregulated genes		Downregulated genes	
Regulator	P-value	Regulator	P-value
RUNX1 (−)	$1.22 \cdot 10^{-180}$	NR2F2 (−)	$7.83 \cdot 10^{-116}$
TCF3 (+)	$5.96 \cdot 10^{-163}$	MAX (+)	$3.27 \cdot 10^{-105}$
NR2F2 (+)	$6.19 \cdot 10^{-163}$	TCF3 (−)	$3.12 \cdot 10^{-95}$
MAX (−)	$3.54 \cdot 10^{-157}$	RUNX1 (+)	$1.78 \cdot 10^{-94}$
SFPQ (+)	$1.06 \cdot 10^{-136}$	CREBBP (−)	$8.51 \cdot 10^{-78}$
ELF1 (−)	$4.60 \cdot 10^{-134}$	ELF1 (+)	$1.09 \cdot 10^{-76}$
KDM5B (+)	$1.68 \cdot 10^{-131}$	SUMO2 (−)	$4.03 \cdot 10^{-74}$
HDAC1 (+)	$9.85 \cdot 10^{-125}$	CREB1 (−)	$4.42 \cdot 10^{-70}$
SIN3A (+)	$2.90 \cdot 10^{-123}$	SMC3 (−)	$8.33 \cdot 10^{-70}$
CREB1 (+)	$5.84 \cdot 10^{-123}$	UBTF (−)	$9.24 \cdot 10^{-61}$
SMC3 (+)	$9.37 \cdot 10^{-120}$	RAD21 (−)	$4.72 \cdot 10^{-65}$
CREBBP (+)	$7.36 \cdot 10^{-119}$	HDAC1 (−)	$1.03 \cdot 10^{-61}$
SUMO2 (+)	$5.37 \cdot 10^{-115}$	SMARCC2 (−)	$3.84 \cdot 10^{-61}$
RAD21 (+)	$7.93 \cdot 10^{-113}$	SFPQ (−)	$5.60 \cdot 10^{-60}$
FOXP1 (−)	$3.66 \cdot 10^{-104}$	FOXP1 (+)	$8.87 \cdot 10^{-59}$
STAT1 (−)	$8.03 \cdot 10^{-104}$	KDM5B (−)	$5.10 \cdot 10^{-56}$
UBTF (+)	$5.37 \cdot 10^{-102}$	STAT1 (+)	$2.86 \cdot 10^{-55}$
ZNF384 (+)	$2.97 \cdot 10^{-101}$	SMAD3 (−)	$1.13 \cdot 10^{-50}$
SMARCC2 (+)	$1.08 \cdot 10^{-94}$	SIN3A (−)	$1.99 \cdot 10^{-50}$
ERG (−)	$2.55 \cdot 10^{-90}$	TAF7 (−)	$1.78 \cdot 10^{-49}$
TAF7 (+)	$6.93 \cdot 10^{-90}$	ZNF384 (−)	$2.22 \cdot 10^{-49}$
SPI1 (−)	$4.17 \cdot 10^{-84}$	SPI1 (+)	$6.10 \cdot 10^{-47}$
HDAC2 (+)	$4.41 \cdot 10^{-82}$	CEBPB (+)	$1.22 \cdot 10^{-46}$
SMAD3 (+)	$1.26 \cdot 10^{-79}$	ERG (+)	$2.65 \cdot 10^{-38}$
HOXA4 (+)	$2.65 \cdot 10^{-75}$	RUNX3 (+)	$9.67 \cdot 10^{-37}$

SIX5 (+)	$3.44 \cdot 10^{-75}$	CTCF (−)	$3.51 \cdot 10^{-35}$
CEBPB (−)	$2.90 \cdot 10^{-70}$	SIX5 (−)	$2.60 \cdot 10^{-33}$
WDR5 (+)	$3.43 \cdot 10^{-66}$	HOXA4 (+)	$5.38 \cdot 10^{-33}$
KDM4A (+)	$1.07 \cdot 10^{-64}$	FOSL1 (+)	$9.67 \cdot 10^{-37}$
BMI1 (+)	$1.94 \cdot 10^{-64}$	STAT5A (+)	$3.04 \cdot 10^{-31}$
SP4 (+)	$2.06 \cdot 10^{-60}$	GABPA (−)	$4.81 \cdot 10^{-31}$
YY1 (+)	$3.37 \cdot 10^{-60}$	BATF (+)	$8.99 \cdot 10^{-31}$
BATF (−)	$1.39 \cdot 10^{-54}$	HDAC2 (−)	$2.48 \cdot 10^{-30}$
CTCF (+)	$2.68 \cdot 10^{-53}$	YY1 (−)	$5.81 \cdot 10^{-30}$
RUNX3 (−)	$3.57 \cdot 10^{-51}$	VDR (+)	$1.56 \cdot 10^{-29}$
STAT5A (−)	$7.98 \cdot 10^{-51}$	NR2F1 (−)	$2.12 \cdot 10^{-29}$
HOXA6 (+)	$1.82 \cdot 10^{-50}$	NFATC1 (+)	$1.25 \cdot 10^{-28}$
MTA3 (+)	$8.70 \cdot 10^{-49}$	IKZF1 (+)	$7.73 \cdot 10^{-28}$
GABPA (+)	$4.30 \cdot 10^{-46}$	FOSL2 (+)	$5.29 \cdot 10^{-26}$
CTBP2 (+)	$8.20 \cdot 10^{-46}$	CTBP2 (−)	$1.48 \cdot 10^{-24}$
SMARCC1 (+)	$3.08 \cdot 10^{-45}$	PPARD (+)	$3.03 \cdot 10^{-24}$
FOSL2 (−)	$2.21 \cdot 10^{-44}$	KDM4A (−)	$4.07 \cdot 10^{-24}$
KLF1 (−)	$6.81 \cdot 10^{-44}$	MTA3 (−)	$9.88 \cdot 10^{-24}$
NFATC1 (−)	$8.68 \cdot 10^{-44}$	EP300 (−)	$1.23 \cdot 10^{-23}$
MAFK (−)	$2.56 \cdot 10^{-43}$	HOXA6 (−)	$2.32 \cdot 10^{-23}$
VDR (−)	$6.54 \cdot 10^{-43}$	KLF1 (+)	$3.00 \cdot 10^{-23}$
EP300 (+)	$7.79 \cdot 10^{-43}$	SP4 (−)	$6.82 \cdot 10^{-23}$
ZBTB33 (+)	$2.59 \cdot 10^{-39}$	MAFK (+)	$1.17 \cdot 10^{-22}$
NR2F1 (+)	$5.83 \cdot 10^{-39}$	WDR5 (−)	$1.80 \cdot 10^{-22}$
DUX4 (−)	$4.33 \cdot 10^{-37}$	IRF4 (+)	$1.38 \cdot 10^{-21}$

Table A.10 Aggregated REGGAE results for upregulated and downregulated genes, respectively. Each ranking was obtained via a sum-of-rank aggregation of the REGGAE results for input lists of the following sizes: 250, 500, 750, and 1,000, as well as all significantly upregulated (538) and downregulated (317) genes (with p-value < 0.01). The colors of the gene symbols in the first and third column indicate whether the mean correlation coefficient between a regulator and its target genes is **positive** (+) or **negative** (−).

The complete list of identified regulators can be accessed via the following link:

Click here to access / download the supplementary file from
www.lara-schneider.de/dissertation/REGGAE_Wilms_Case_study_full_results.xlsx

A.6 Supplements for NetworkTrail

The NetworkTrail web service enables users to detect the most deregulated pathways and subgraphs in biological networks (cf. **Section 4.4**). The resulting subgraphs can be downloaded in `SIF` and `NA` format, which can then be visualized in a variety of network visualization tools. Details on these file formats are provided in the following section.

A.6.1 SIF and NA file formats

The Simple Interaction Format (`SIF`) is a text-based format that builds graphs from a list of interactions. Each line in the file describes an interaction, consisting of a source node, the edge type (e.g., 'activation' or 'inhibition'), and one or more target nodes:

```
node1 typeA node2
node2 typeB node3 node4
node3 typeA node4
```

In order to provide additional information about the nodes in a graph, Node Attribute (`NA`) files can be used. An `NA` file begins with the name of the attribute in the first line. Each of the remaining lines contains the identifier of a node followed by '`=`' and the value of that attribute. By this, one can, for example, assign weights to the network nodes:

```
nodeWeights
node1 = 0.82
node2 = -1.3
node3 = -0.42
node4 = 2.35
```

A.7 Supplements for DrugTargetInspector

DrugTargetInspector (DTI) is an interactive assistance tool that provides rich functionality for the integrative analysis of tumor-specific genomics, transcriptomics, and proteomics data sets (cf. **Chapter 5**).

A.7.1 Provided functionality

The main results page of DTI provides a variety of options for the in-depth analysis of the uploaded data sets. Many of those can be accessed via the side panel. **Figure A.5** provides an overview of the side panel content. One of the in-depth analyses provided here is the subgraph analysis, in which the most

deregulated subnetworks rooted in drug targets of interest are computed based on the KEGG regulatory signaling network (cf. **Section 5.2.3.5**). To this end, an Integer Linear Programming formulation is used, which is based on the 'Subgraph ILP' described in **Section 3.3.3.3** and **Table 3.1**. In its adapted formulation for the use in DrugTargetInspector, a specific drug target t is fixed to be the root node, see **Table A.11** for the complete formulation.

Objective		
$\max_{x \in \mathbb{B}^n} \sum_i w_i \cdot x_i$	(A.1)	Maximize the overall deregulation of the subgraph
Subject to		
$\sum_i x_i = k$	(A.2)	Ensures that the subgraph is of size k
$\sum_i y_i = 1$	(A.3)	Ensures that a single root node is selected
$y_t = 1$	(A.4)	Ensures that the considered molecular drug target t is the root node
$y_i \leq x_i \quad \forall i$	(A.5)	Ensures that the designated root node is part of the selected subgraph
$x_i - y_i - \sum_{j \in In(i)} x_j \leq 0 \quad \forall i$	(A.6)	Ensures connectivity of the subgraph
$\sum_{i \in C} (x_i - y_i) - \sum_{j \in In(C)} x_j \leq \lvert C \rvert - 1 \quad \forall C$	(A.7)	Prevents disconnected cycles

Table A.11 **Subgraph ILP formulation**. The objective function and the respective constraints are given in the first column, the second column provides a numbering for reference in the text, and the third column describes the purpose of the respective formula. The variable C describes cycles formed by the selected nodes.

A.7.2 Wilms tumor data set

In one of the presented case studies for DrugTargetInspector, we investigated Wilms tumor samples of several subtypes and analyzed their transcriptomic profiles to identify deregulated drug targets and altered biological pathways that might inform the selection of adjuvant treatment options (cf. **Section 5.3.1**). For our analysis, we used a gene expression data set of 37 Wilms tumor samples of different subtypes (cf. **Table A.12**).

A.8 Supplements for ClinOmicsTrail^{bc}

ClinOmicsTrail^{bc} is a comprehensive visual analytics tool for breast cancer decision support that provides a holistic assessment of standard-of-care targeted drugs, candidates for drug repositioning, and immunotherapeutic approaches (cf. **Chapter 6**). In the following sections, we provide supplementary information on ClinOmicsTrail^{bc}'s functionality and additional results for the case studies discussed in **Section 6.5**.

Figure A.5 Overview of side panel content on DrugTargetInspector's results page. The side panel on DTI's results page (middle) consists of eleven foldable sub-panels, which unfold when being clicked on. The content of the respective unfolded sub-panels is shown with color-coordinated borders matching their folded counterparts.

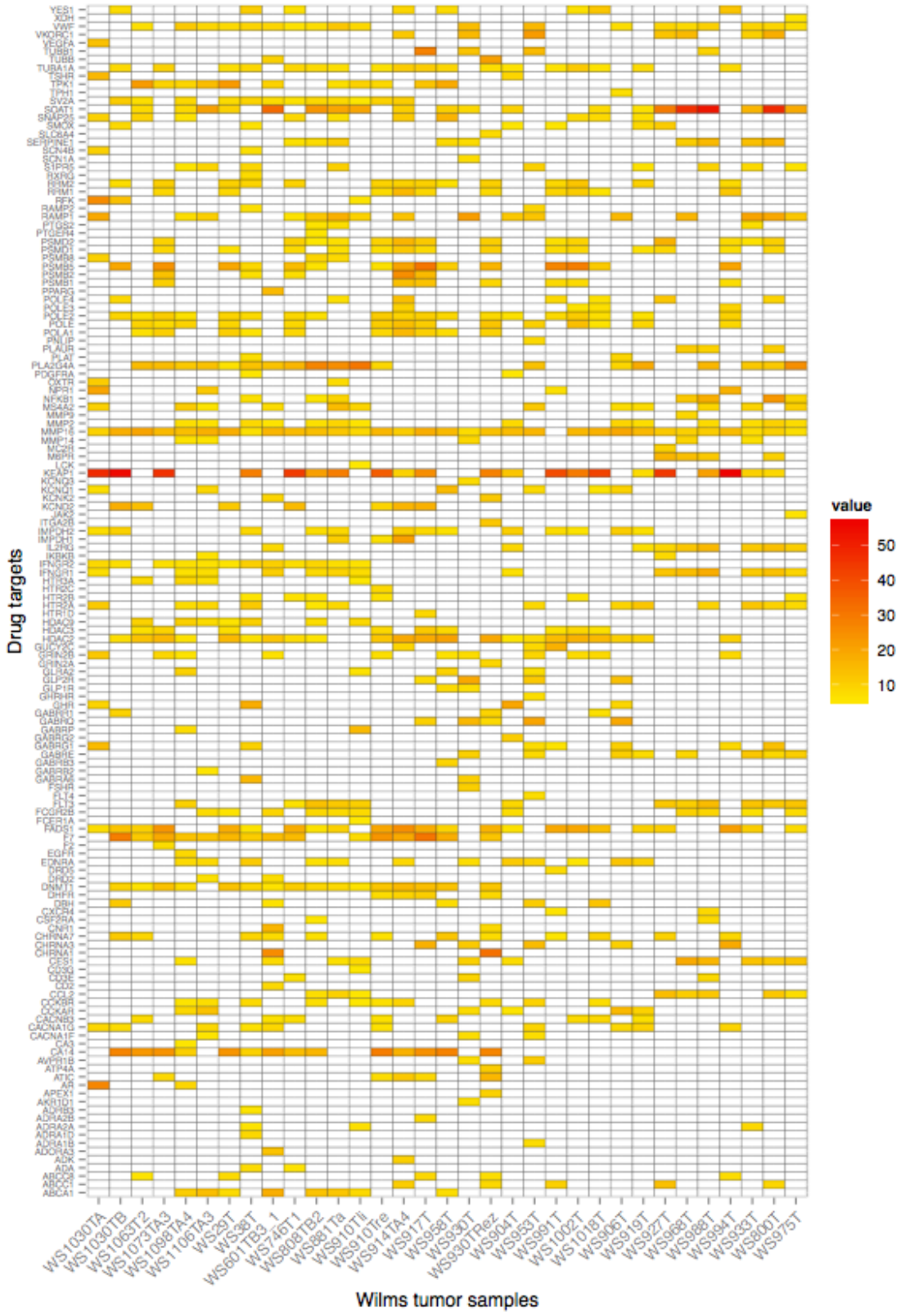


Figure A.6 Wilms tumor samples' expression of drug targets. Heat map of Wilms tumor samples and a consensus set of deregulated drug targets in DrugTargetInspector. Colored cells in the heat map correspond to significantly upregulated drug targets in the respective samples. White cells indicate that the corresponding drug target was not significantly upregulated in the respective sample.

Histologic phenotype	Samples
Blastema (high risk)	**WS1030TA**, **WS1030TB**, **WS1063T2**, **WS1073TA3**, **WS1098TA4**, **WS1106TA3**, **WS29T**, **WS38T**, **WS601TB3_1**, **WS746T1**, **WS808TB2**, **WS881Ta**, **WS910Tli**, **WS910Tre**, **WS914TA4**, **WS917T**, **WS958T**
Diffuse anaplasia (high risk)	**WS930T**, **WS930Trez**
Stromal type (intermediate risk)	**WS904T**
Focal anaplasia (intermediate risk)	**WS953T**
Epithelial type (intermediate risk)	**WS991T**
Triphasic (intermediate risk)	**WS1002T**, **WS1018T(re)**, **WS906T**, **WS919T**
Regressive (intermediate risk)	**WS927T**, **WS968T**, **WS988T**, **WS994T**, **WS933T**
Completely necrotic (low risk)	**WS800T**, **WS975T**
Normal	WS1018Ni_1, WS1018Ni_2, WS968Ni, WS878Ni

Table A.12 **Wilms tumor data set**. Histologic phenotypes and associated risk of relapse for Wilms tumor samples used in DrugTargetInspector case study. For the analysis using REGGAE (cf. **Section 4.3.4.2**), the samples printed in **bold** were used.

A.8.1 Pathway activity measure

ClinOmicsTrail^{bc} offers functionality to assess pathway activities for a set of 20 core breast cancer-relevant pathways (cf. **Section 6.4.2.2**). The gene sets representing each of those pathways are discussed in **Section A.8.1.1** and an overview of the pathway activity distribution over different molecular breast cancer subtypes is shown in **Section A.8.1.2**.

A.8.1.1 Metapathways for pathway activity computation

For the pathway activity computation approach described in **Section A.8.1.2**, we consider a set of 20 'metapathways'. These metapathways are gene sets that we obtain by taking the union of relevant gene sets from KEGG [282], GO [424], Reactome [283], and WikiPathways [628]. For an overview of the considered gene sets and their assignment to the metapathways, please download the supplementary file linked below:

Click here to access / download the supplementary file from
www.lara-schneider.de/dissertation/ClinOmicsTrail_Metapathways.xlsx

A.8.1.2 Computation of pathway activity measure

Figure A.7 gives an overview of computed pathway activities for the TCGA breast cancer cohort, grouped according to their pathological subtypes.

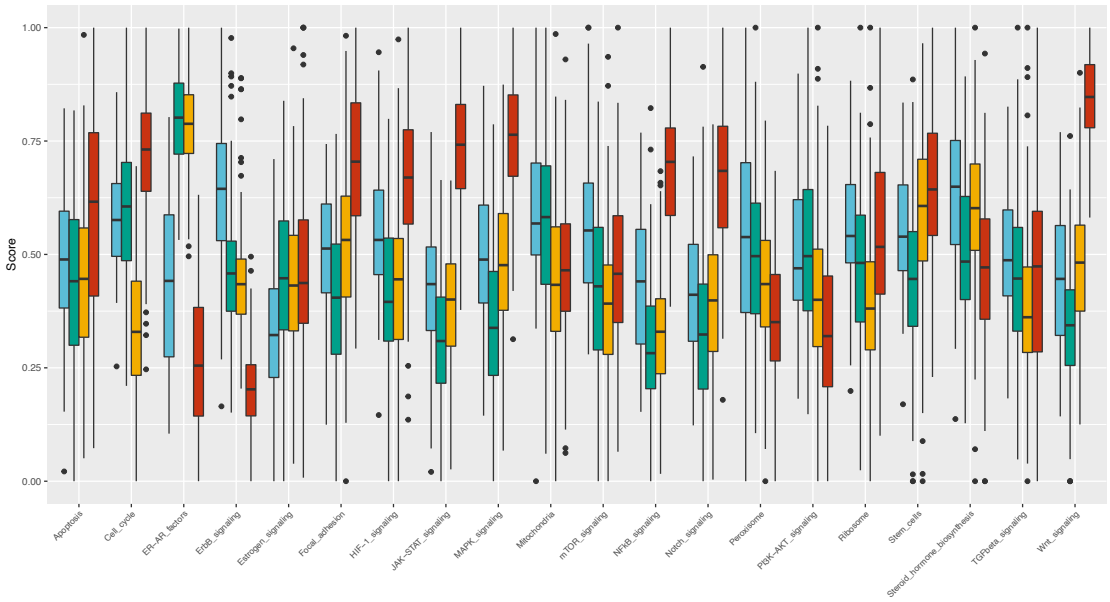


Figure A.7 Subtype-specific pathway activities. Boxplot for pathway activities TCGA breast cancer samples. Scores of differential gene expression were computed as z-scores comparing a tumor sample against the cohort of normal samples. Samples color-coded based on subtype given in data set. **Blue**: HER2-enriched, **green**: luminal B, **orange**: luminal A, **red**: basal-like.

A.8.2 Rule-based drug assessment

ClinOmicsTrail^{bc} performs an assessment of a variety of standard-of-care drugs by considering several classes of genes, proteins, and pathways that might promote or hinder the effectiveness of a drug. For a set of 17 FDA-approved, standard-of-care breast cancer drugs (cf. **Figure 6.7**), ClinOmicsTrail^{bc} assesses the genomic and transcriptomic status of respective molecular drug targets, drug-processing enzymes, resistance-promoting factors, and associated pathways. Since the respective categories reflect different mechanisms that might (de)sensitize a tumor with regard to the considered drug, different clinical, genomic, and transcriptomic traits have to be considered in each case. The following figures provide an overview of the rule-based assessment applied for biomarkers (**Figure A.8**), drug targets (**Figure A.9**), ADME genes (**Figure A.10**), transporters (**Figure A.11**), and pathways (**Figure A.12**).

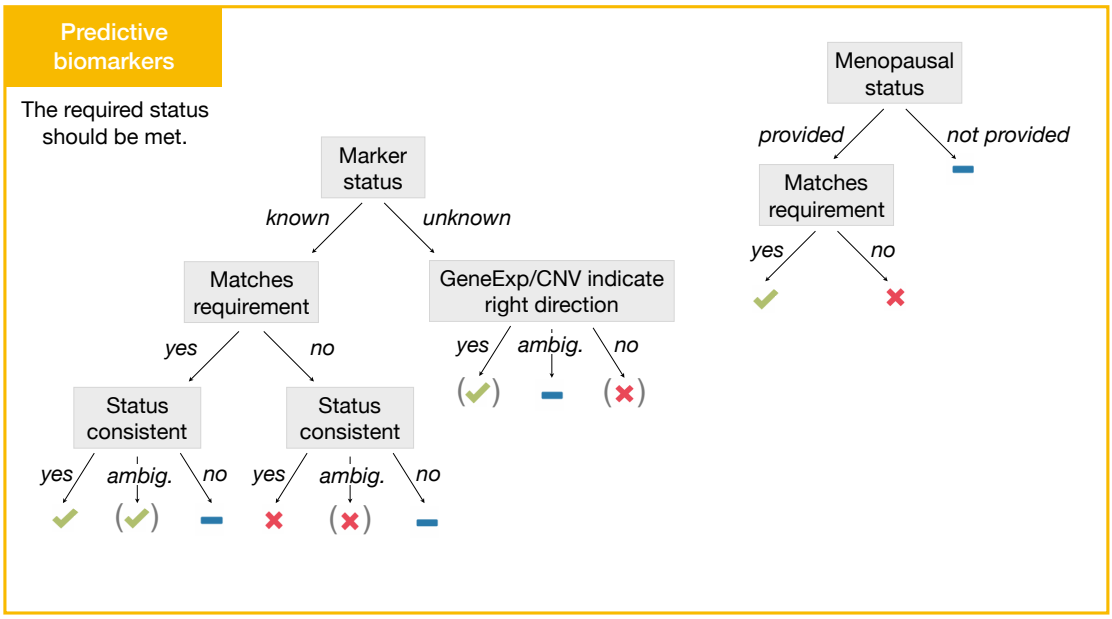


Figure A.8 **Biomarker evaluation for rule-based drug assessment**. The green checkmark symbol indicates that there seems to be no impediment for the efficacy of the drug of interest with respect to the considered biomarker. The blue minus symbol denotes that there might be some impediments and the red cross symbol highlights that there seem to be contraindications to the successful treatment with the considered drug. Symbols in parentheses mean that there are inconsistencies in the data.

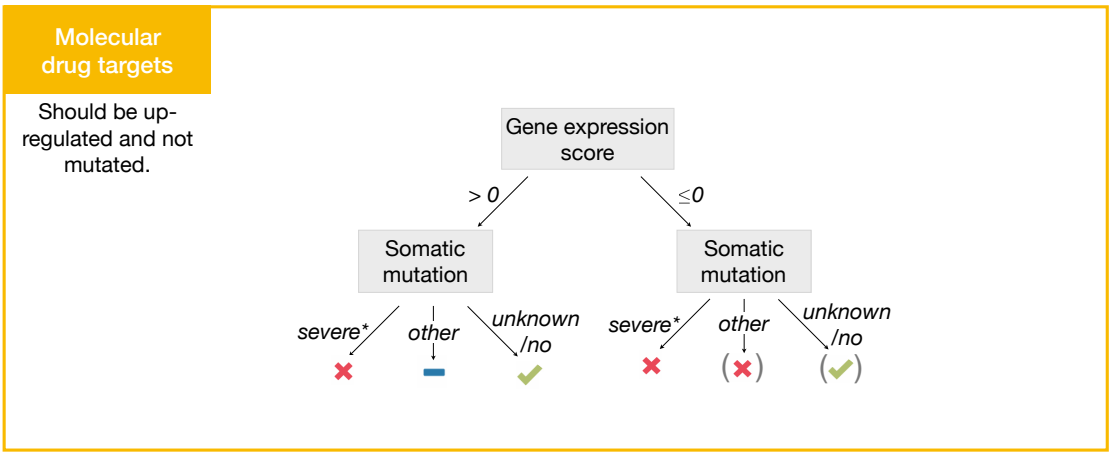


Figure A.9 **Drug target evaluation for rule-based drug assessment**. The green checkmark symbol indicates that there seems to be no impediment for the efficacy of the drug of interest with respect to the considered drug target. The blue minus symbol denotes that there might be some impediments and the red cross symbol highlights that there seem to be contraindications to the successful treatment with the considered drug. Symbols in parentheses mean that there are inconsistencies in the data.
*Severe mutations = {frameshift, stop lost, stop gained, start lost}

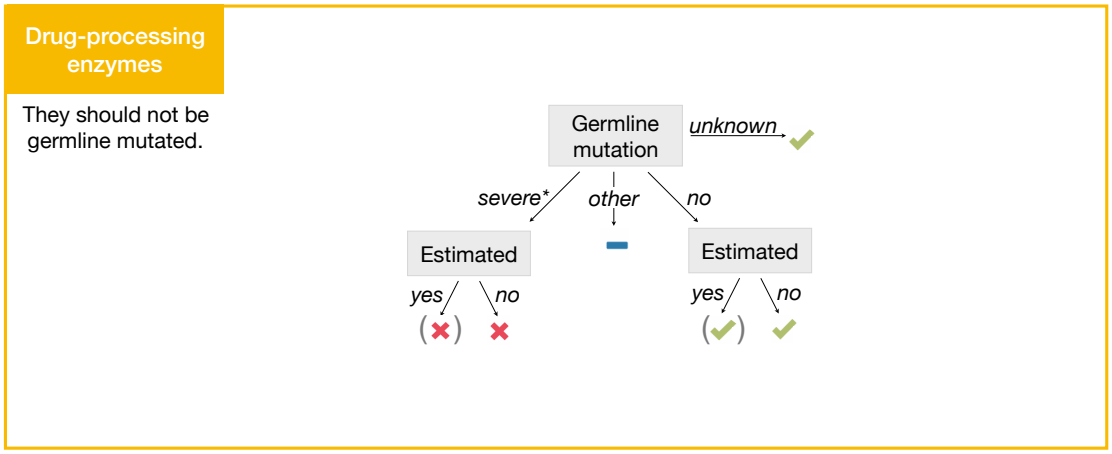


Figure A.10 **Drug-processing enzyme evaluation for rule-based drug assessment**. The green checkmark symbol indicates that there seems to be no impediment for the efficacy of the drug of interest with respect to the considered drug-processing enzyme. The blue minus symbol denotes that there might be some impediments and the red cross symbol highlights that there seem to be contraindications to the successful treatment with the considered drug. Symbols in parentheses mean that there are inconsistencies in the data.
*Severe mutations = {frameshift, stop lost, stop gained, start lost}

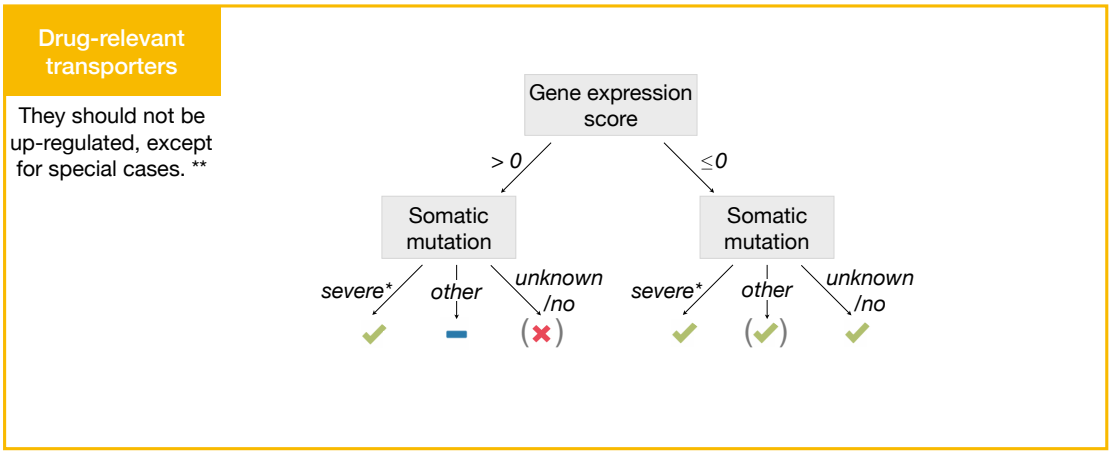


Figure A.11 **Transporter evaluation for rule-based drug assessment**. The green checkmark symbol indicates that there seems to be no impediment for the efficacy of the drug of interest with respect to the considered efflux transporter. The blue minus symbol denotes that there might be some impediments and the red cross symbol highlights that there seem to be contraindications to the successful treatment with the considered drug. Symbols in parentheses mean that there are inconsistencies in the data.
*Severe mutations = {frameshift, stop lost, stop gained, start lost}
**Special cases: tamoxifen, lapatinib, and abemaciclib that inhibit certain transporters

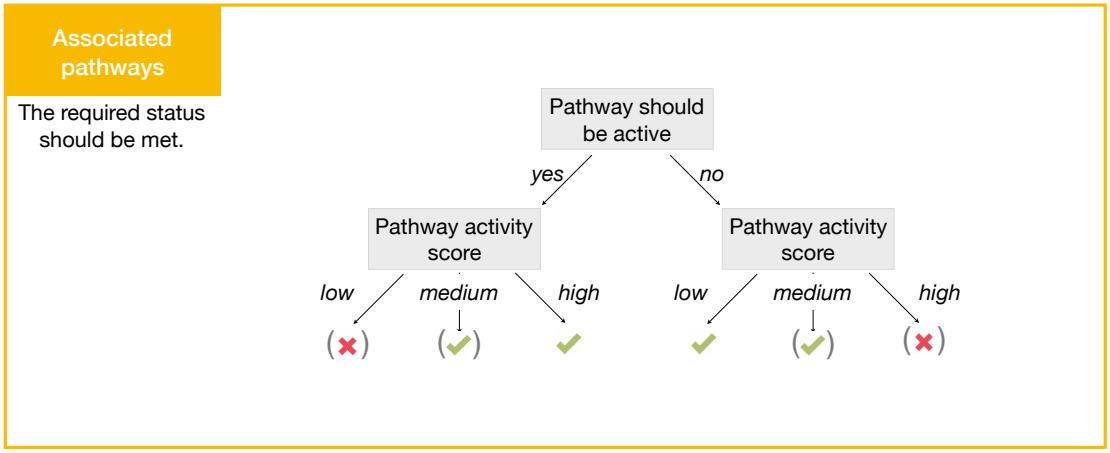


Figure A.12 **Pathway activity assessment for rule-based drug assessment**. The green checkmark symbol indicates that there seems to be no impediment for the efficacy of the drug of interest with respect to the considered pathway's activity. The blue minus symbol denotes that there might be some impediments and the red cross symbol highlights that there seem to be contraindications to the successful treatment with the considered drug. Symbols in parentheses mean that there are inconsistencies in the data. *Low*: pathway activitiy score in $[0, 0.4)$, *medium*: pathway activity score in $[0.4, 0.6]$, *high*: pathway activity score in $(0.6, 1]$.

A.8.3 Deficient repair machinery and tumor mutational burden

A high tumor mutational load in a cancer sample is likely to be fostered by deficiencies in the DNA repair machinery [6]. To further underline the connection between TMB and an impaired DNA repair machinery we compared the 100 TCGA samples with the highest TMB to the 100 samples with the lowest TMB and tested for enrichment in a set of 52 repair-related biological categories (cf. **Section 6.5.3**). Additional details on the considered gene sets and their originating databases can be found in this supplementary file:

Click here to access / download the supplementary file from
www.lara-schneider.de/dissertation/ClinOmicsTrail_Repair_gene_enrichment.xlsx

Table A.13 lists the 32 significantly enriched categories.

Name	Number of hits	Expected score	Adjusted p-value
GO Biological Process nucleotide excision repair (5)	62	14.9813	1.02E-13
GO Biological Process regulation of DNA repair (5)	46	10.2276	3.40E-11
Reactome SUMOylation of DNA damage response and repair proteins	43	9.6514	1.67E-10
GO Biological Process non recombinational repair (5)	34	7.20254	7.96E-9
GO Biological Process double strand break repair via nonhomologous end joining (6)	31	6.48229	3.25E-8
GO Biological Process global genome nucleotide excision repair (6)	31	7.49064	3.42E-7

GO Biological Process postreplication repair (5)	30	7.20254	4.54E-7
GO Biological Process double strand break repair via synthesis dependent strand annealing (7)	22	3.74532	4.55E-7
GO Biological Process base excision repair (5)	26	5.90608	1.42E-6
KEGG Nucleotide excision repair	25	5.90608	3.99E-6
GO Biological Process regulation of double strand break repair (6)	22	4.60962	4.04E-6
GO Biological Process transcription coupled nucleotide excision repair (6)	32	9.50735	4.04E-6
GO Cellular Component DNA repair complex (4)	23	5.18583	5.16E-6
GO Biological Process positive regulation of DNA repair (5)	21	5.04178	3.29E-5
GO Biological Process mismatch repair (5)	19	4.32152	5.23E-5
GO Biological Process nucleotide excision repair DNA incision (6)	19	4.75368	1.36E-4
Reactome Gap-filling DNA repair synthesis and ligation in TC-NER	25	7.92279	1.37E-4
KEGG Base excision repair	17	4.17747	2.98E-4
Reactome Recruitment and ATM-mediated phosphorylation of repair and signaling proteins at DNA double strand breaks	27	9.9395	4.65E-4
KEGG Mismatch repair	13	2.59291	5.24E-4
GO Biological Process interstrand cross link repair (5)	17	4.89773	1.16E-3
GO Biological Process DNA synthesis involved in DNA repair (5)	10	1.87266	2.55E-3
GO Biological Process regulation of double strand break repair via homologous recombination (7)	10	2.01671	3.71E-3
WikiPathways Mismatch repair	7	0.864305	4.21E-3
Reactome Transcription-Coupled Nucleotide Excision Repair (TC-NER)	17	5.90608	5.48E-3
GO Biological Process nucleotide excision repair DNA gap filling (6)	10	2.30481	6.75E-3
GO Biological Process positive regulation of double strand break repair (6)	7	1.00836	6.75E-3
Reactome Mismatch repair (MMR) directed by MSH2:MSH3 (MutSbeta)	8	1.44051	6.92E-3
Reactome Mismatch repair (MMR) directed by MSH2:MSH6 (MutSalpha)	8	1.44051	6.92E-3

GO Biological Process nucleotide excision repair DNA damage recognition (6)	11	3.02507	9.71E-3
GO Cellular Component mismatch repair complex (4)	7	1.15241	9.71E-3
Reactome Gap-filling DNA repair synthesis and ligation in GG-NER	9	2.30481	1.78E-2

Table A.13 Significantly enriched gene sets from GeneTrail2 Over-Representation Analysis of 52 DNA and mismatch repair machinery-related gene sets. The database a respective gene set originated from is given as a prefix to the name of the gene set in the second column. P-values were FDR-adjusted to a significance level of 0.05.

A.8.4 Neoepitope prediction

Besides checkpoint blockade, personalized cancer vaccines are another promising approach to cancer immunotherapy [88, 89]. Cancer vaccines target overexpressed or altered proteins and HLA presented peptides sequences (neoepitopes) that resulted from somatic mutations uniquely characterizing the patient's tumor. They are used to prime T cells to recognize these characterizing antigens and destroy the presenting tumor cells. As the neoepitopes are dependent on both the patient's tumor mutations and HLA genotype, cancer vaccines have to be individually designed. Thus, ClinOmicsTrail^{bc} offers functionalities to predict potential neoepitope vaccine targets based on the identified somatic mutations and HLA genotype of a patient using the immunoinformatic toolbox ImmunoNodes [647]. ImmunoNodes provides various classes of epitope prediction methods to compute (neo-)epitopes and to assess their affinity to the patient's set of HLA alleles. Details on the 13 methods for neoepitope prediction provided by ClinOmicsTrail^{bc} are listed in **Table A.14**.

Method	Version	Class	Reference
ARB	1.0	MHC-I binding	[716]
BIMAS	1.0	MHC-I binding	[717]
Comblib 2008	1.0	MHC-I binding	[718]
NetMHC	4.0	MHC-I binding	[684]
NetMHCII	2.2	MHC-II binding	[719]
NetMHCIIpan	3.1	MHC-II binding	[720]
NetMHCpan	3.0	MHC-I binding	[719]
PickPocket	1.1	MHC-I binding	[721]
SMM	1.0	MHC-I binding	[722]
SMMPMBEC	1.0	MHC-I binding	[723]
SVMHC	1.0	MHC-I binding	[724]
SYFPEITHI	1.0	T-cell epitope	[725]
UniTope	1.0	T-cell epitope	[726]

Table A.14 Neoepitope prediction method provided by ClinOmicsTrail^{bc}. This table contains the 13 neoepitope prediction methods from ImmunoNodes employed in ClinOmicsTrail^{bc}. The first column contains the tools' names in alphabetical order, the second column the respective version number, the third column the class of the prediction, and the last column the references to the corresponding publications. Table adapted from [647].

A.8.5 Case studies

The following sections contain additional information and results for the analysis of primary tumor samples for the three exemplary TCGA breast cancer samples discussed in **Section 6.5**.

A.8.5.1 Case Study I: TCGA-AN-A0XN

TCGA sample of a 68-year-old (presumably postmenopausal) woman with stage III breast cancer of TNM stage T2/N2/M0. The ER status is negative, PR is positive, and HER2 is not amplified. The tumor sample was predicted to be of *luminal A* subtype by the PAM50 classifier [727].

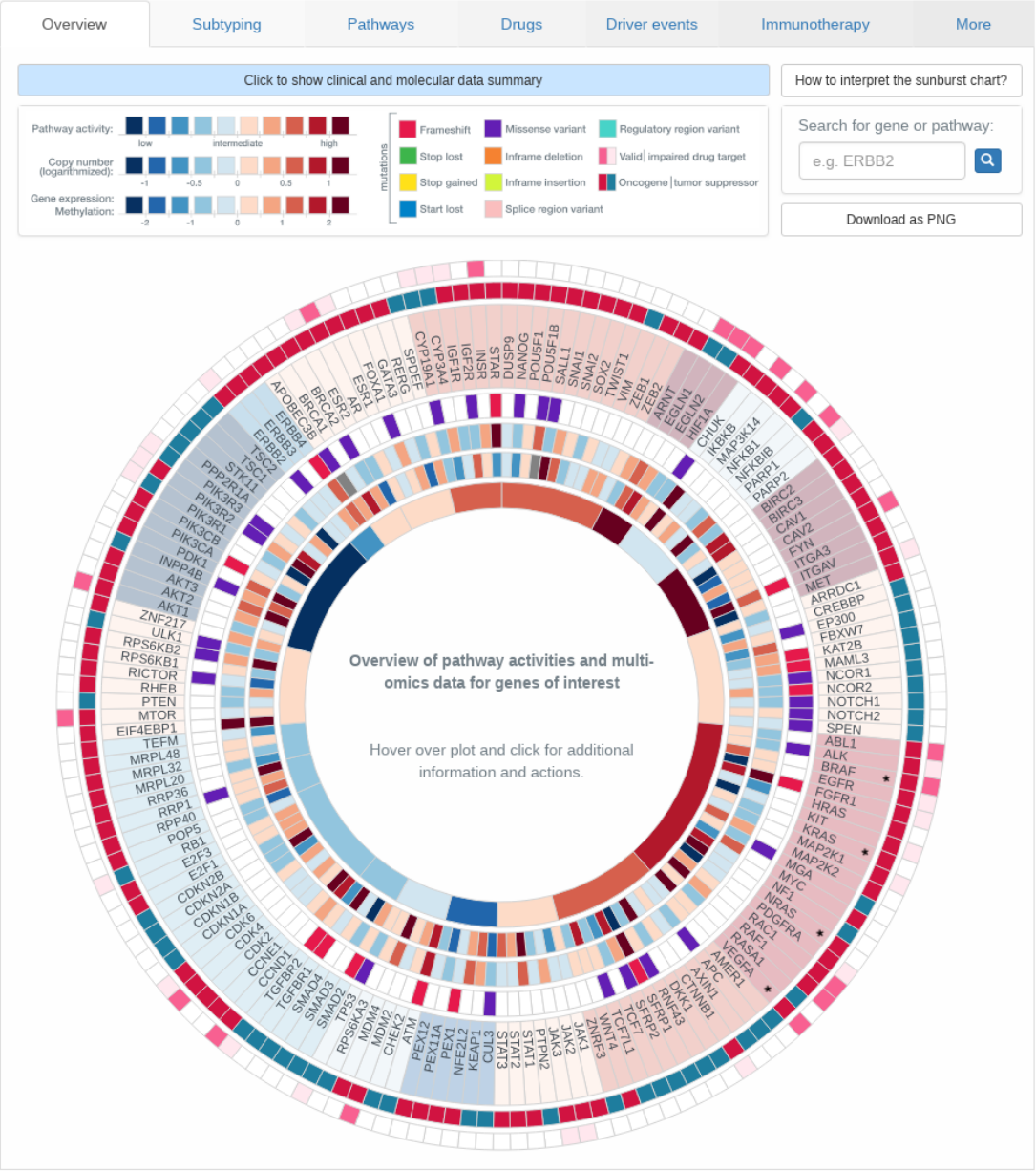


Figure A.13 Sunburst chart overview for TCGA-AN-A0XN. Breast cancer-relevant driver genes and pathways are displayed in a circular manner. Genes are grouped according to the pathways they are most characteristic for. The plot is organized in rings, where the innermost ring displays pathway activities, the second 'inner' ring corresponds to gene expression. Depending on the data provided by the user, information on copy number alterations, and mutations is shown in the third and fourth ring, respectively. Gene names are displayed in the next ring. The second most outer ring indicates whether the gene acts as an oncogene or tumor suppressor gene (TSG) for activating the corresponding pathway. The outermost ring contains indicators on whether or not the gene is a known drug target. Genes discussed in the manuscript are highlighted with an asterisk.

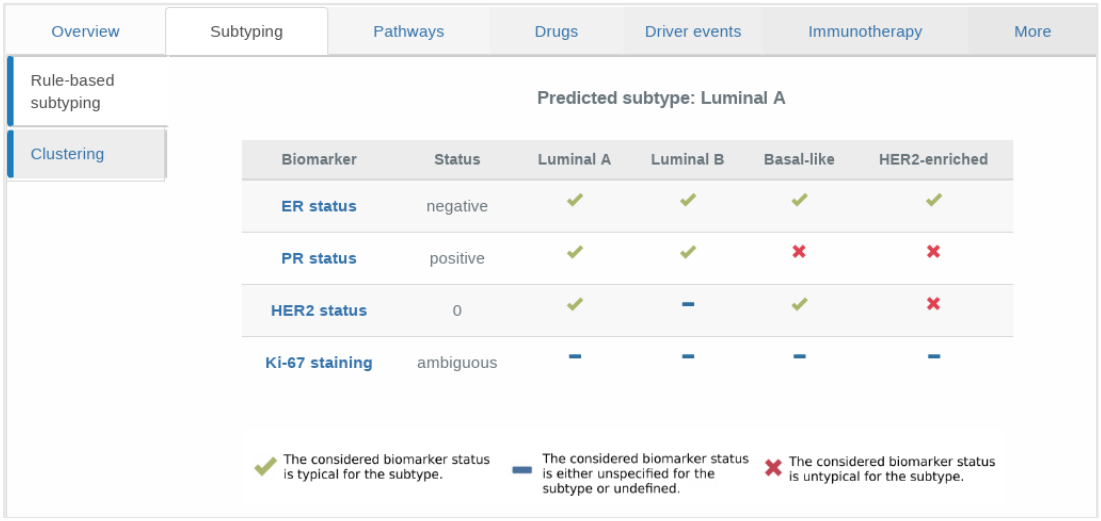


Figure A.14 **Rule-based subtyping for TCGA-AN-A0XN**. Based on the hormone receptor and HER2 status of a tumor sample, as well as the observed growth rates, a classification into the four main breast cancer subtypes luminal A, luminal B, basal-like, and HER2-enriched can be performed.

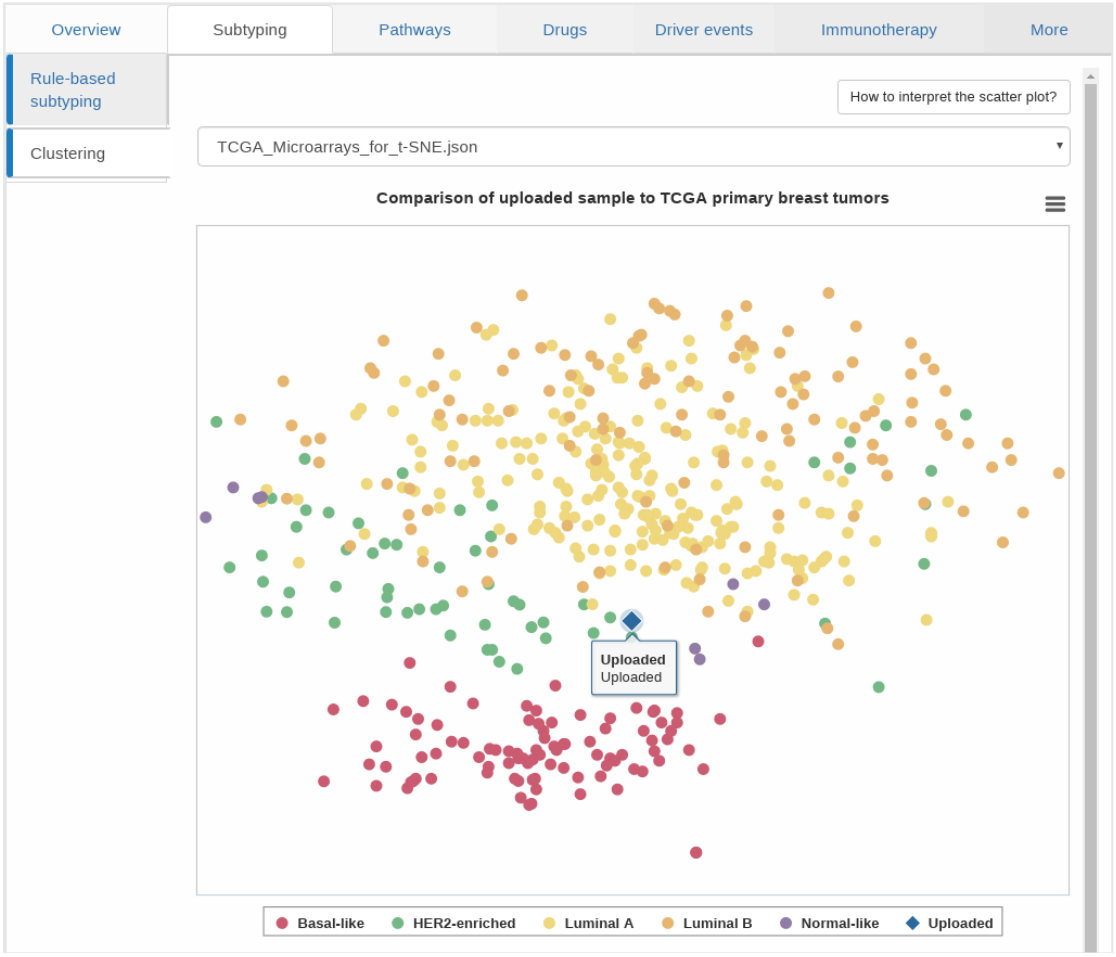


Figure A.15 **Clustering results for TCGA-AN-A0XN**. The tumor sample of interest is clustered along with primary breast tumor samples from TCGA. The molecular subtypes of the TCGA samples are color-coded as indicated by the legend below the plot. The tumor sample under investigation is indicated by the blue diamond-shaped symbol.

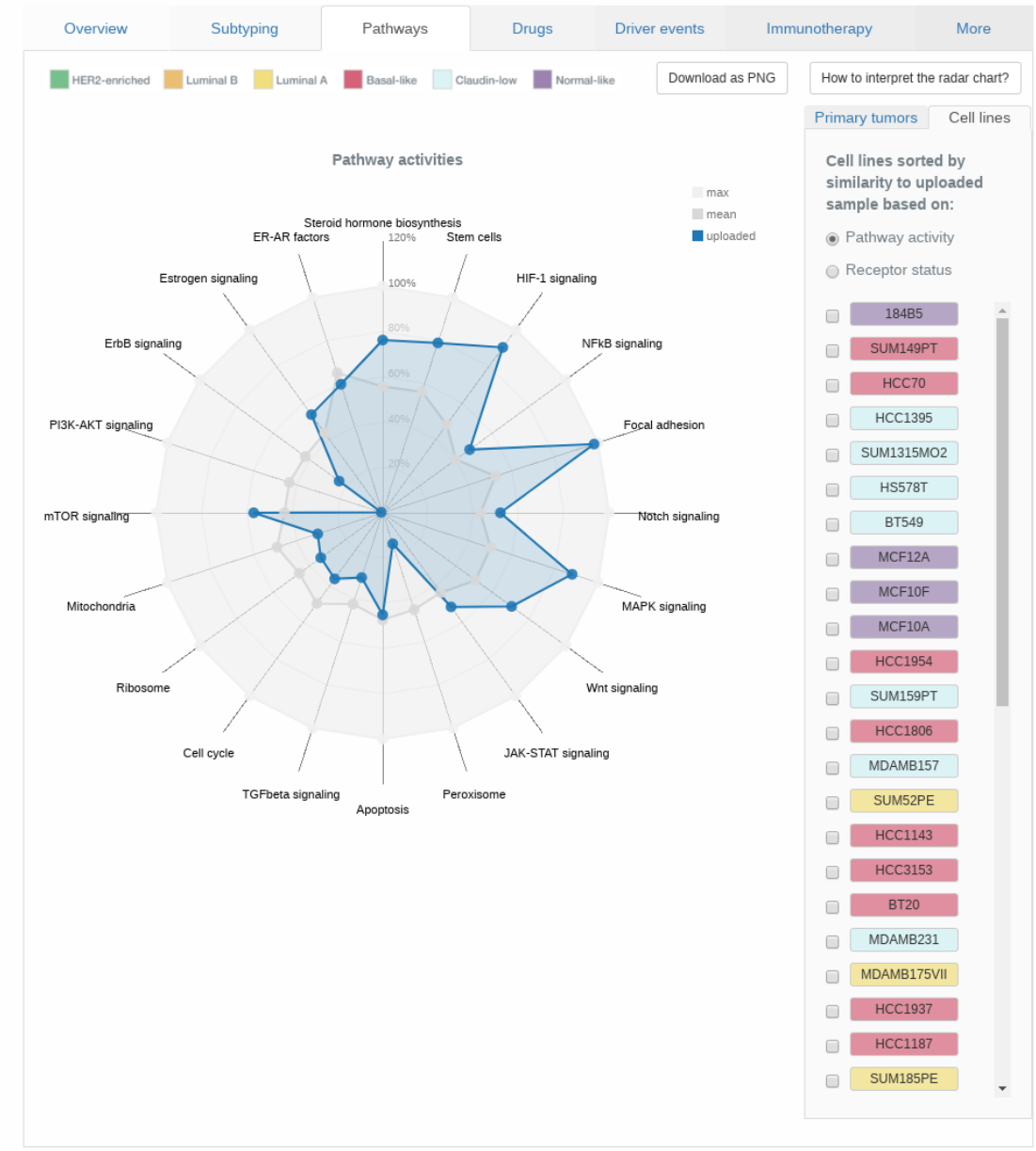


Figure A.16 **Radar chart of pathway activities for TCGA-AN-A0XN**. The pathway activities of a set of 20 core breast cancer pathways for the user-provided tumor sample colored in blue. Reference samples from TCGA as well as breast cancer cell lines can be added to the visualization interactively. The molecular subtype of the respective reference samples is color-coded: basal-like - red, claudin-low - light blue, HER2-enriched - green, luminal A - yellow, luminal B – orange, normal-like - purple. Clicking on a reference sample's name yields additional clinical and pharmacological information, see **Figure A.17**.

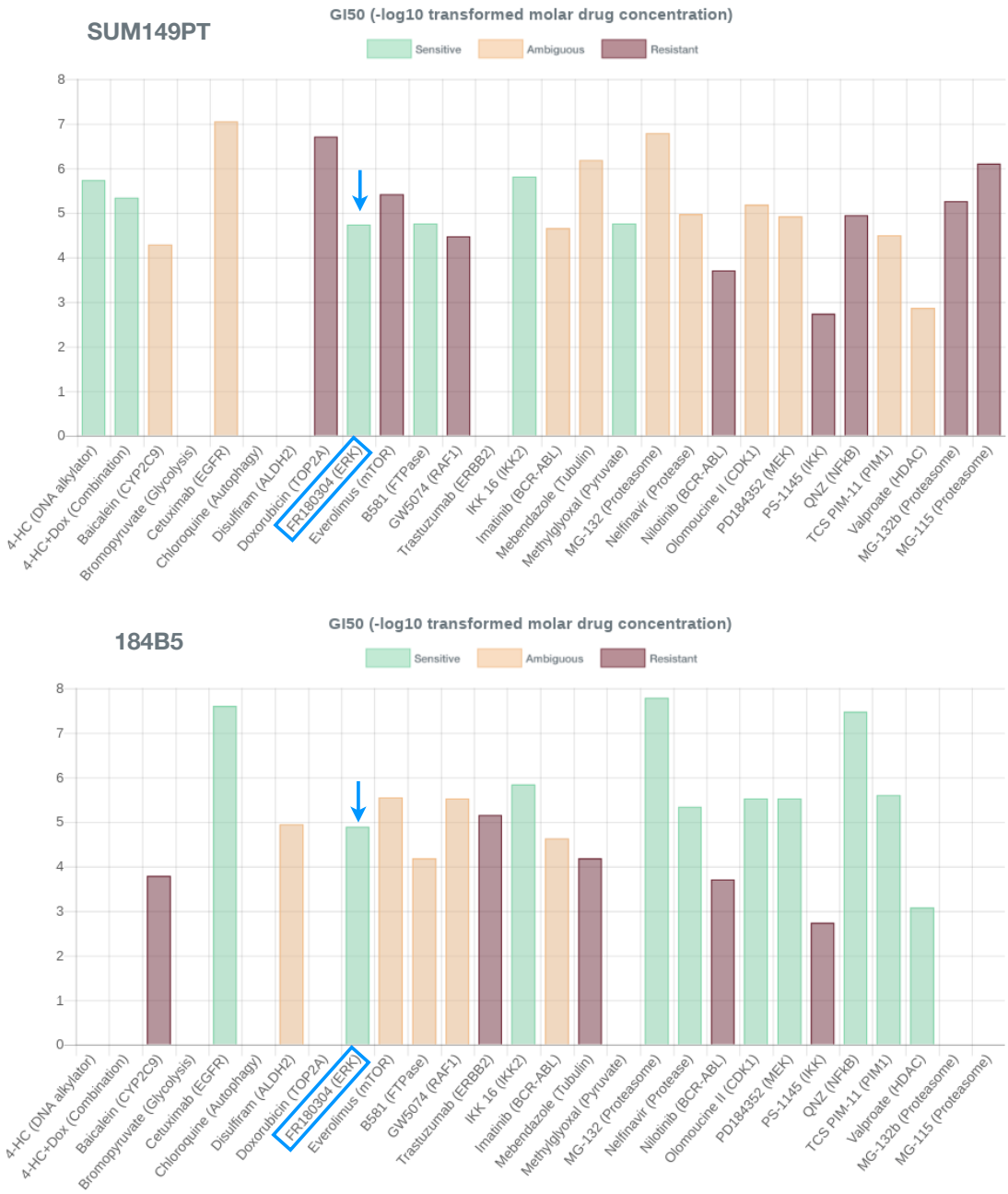


Figure A.17 **Drug sensitivity information for cell lines similar to TCGA-AN-A0XN**. Similarity was assessed based on similarity of pathway activity patterns. The triple negative cell lines SUM149PT and 184B5 are most similar to the sample under investigation **Figure A.16**. Especially, both cell lines were tested to be sensitive for ERK inhibitors (highlighted in blue) by Heiser *et al.* [489].

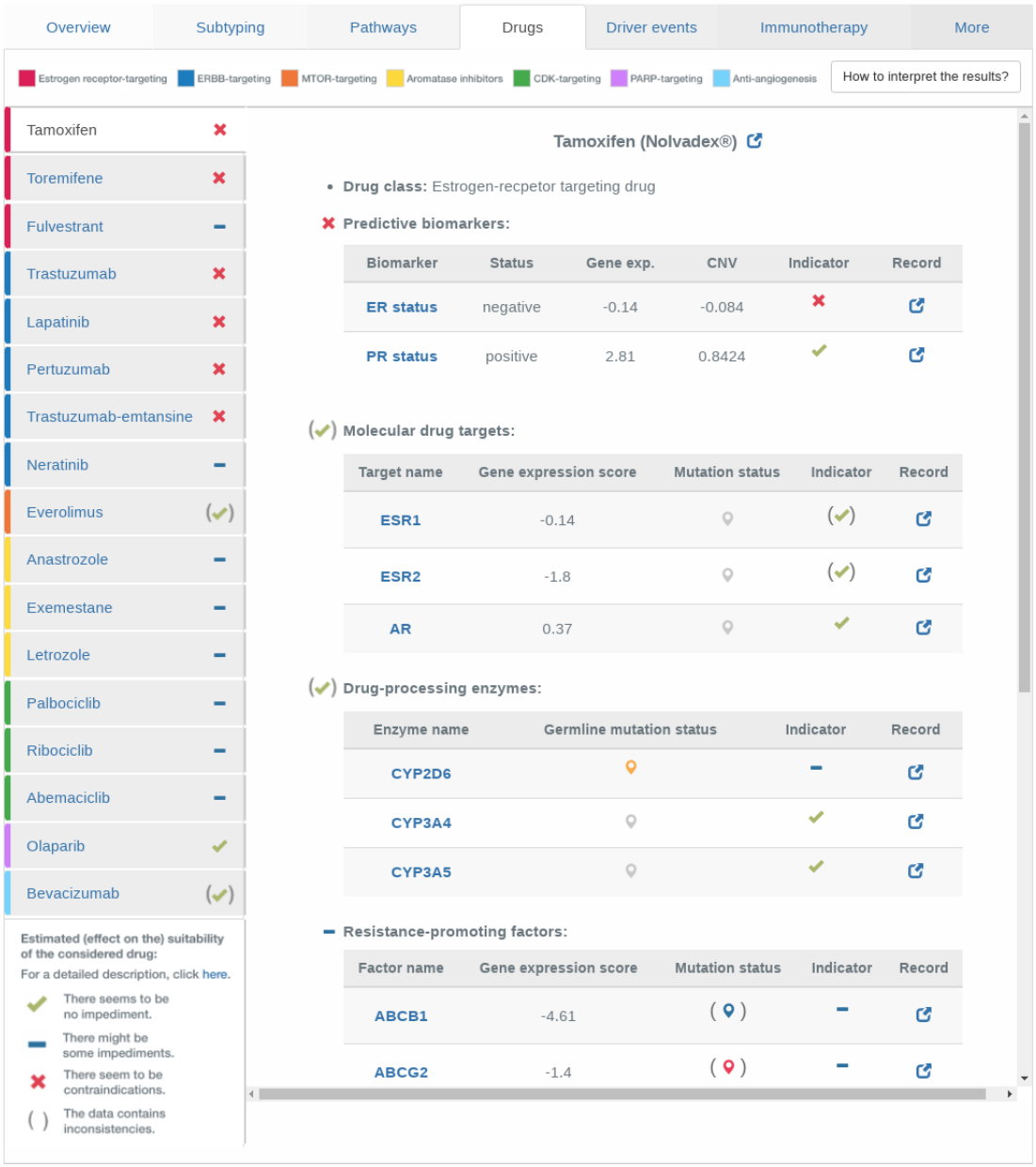


Figure A.18 Assessment of standard-of-care drugs for sample TCGA-AN-A0XN. For a set of 17 standard-of-care breast cancer drugs (left panel), various factors increasing or decreasing the efficacy of a drug are assessed. Clinical, genetic, and molecular characteristics are listed with an indicator sign on whether they might decrease efficacy or even cause resistance to the treatment with the drug under consideration. All genes and pathways are linked to third-party resources where additional details can be found. Each entry also contains the link to a record or publication that describes the role of the corresponding gene with respect to the drug of interest.

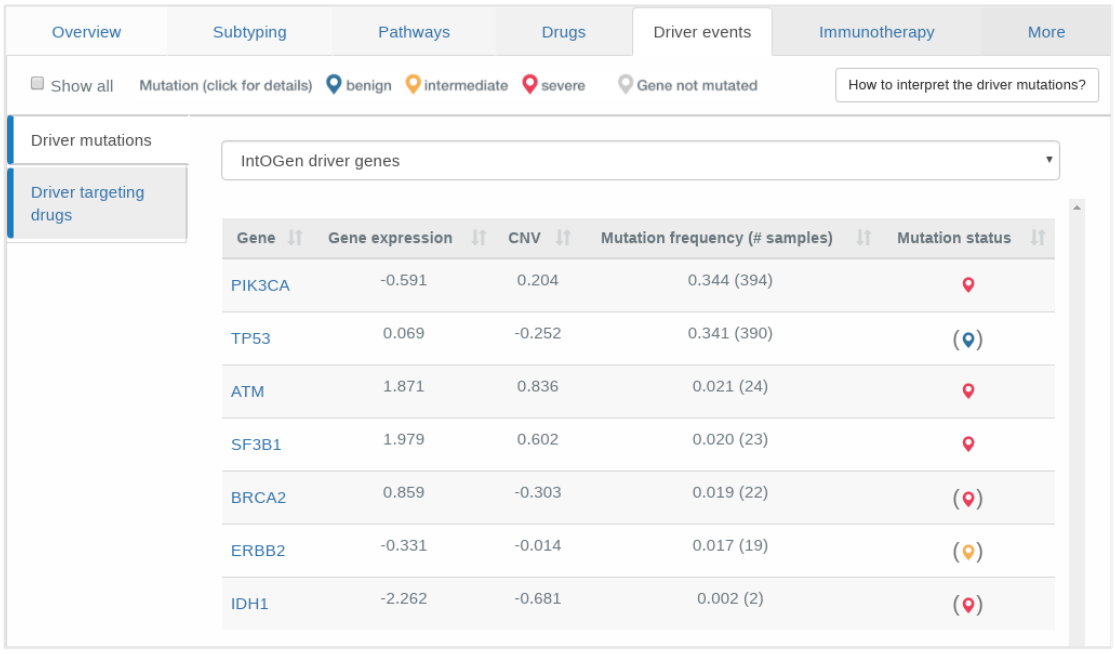
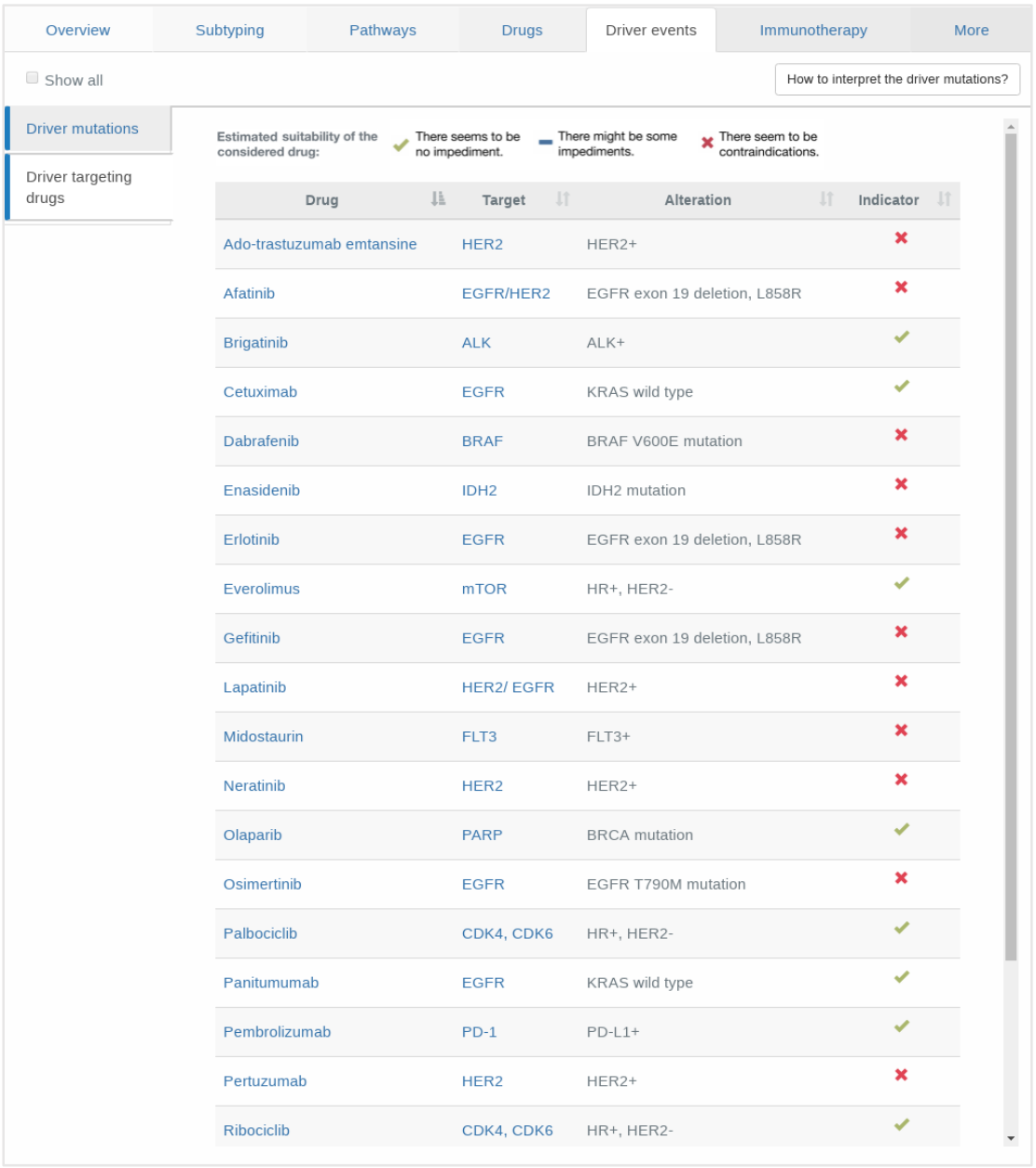


Figure A.19 Driver mutations in sample TCGA-AN-A0XN. Driver mutations contained in the sample under investigation. The color-code in the *Mutation* column indicates the severity of the mutation. Parentheses indicate a germline mutation. Clicking on the indicator symbol opens a modal with additional details on the specific contained mutation(s) and their predicted effect on protein functionality.



Drug	Target	Alteration	Indicator
Ado-trastuzumab emtansine	HER2	HER2+	✗
Afatinib	EGFR/HER2	EGFR exon 19 deletion, L858R	✗
Brigatinib	ALK	ALK+	✓
Cetuximab	EGFR	KRAS wild type	✓
Dabrafenib	BRAF	BRAF V600E mutation	✗
Enasidenib	IDH2	IDH2 mutation	✗
Erlotinib	EGFR	EGFR exon 19 deletion, L858R	✗
Everolimus	mTOR	HR+, HER2-	✓
Gefitinib	EGFR	EGFR exon 19 deletion, L858R	✗
Lapatinib	HER2/ EGFR	HER2+	✗
Midostaurin	FLT3	FLT3+	✗
Neratinib	HER2	HER2+	✗
Olaparib	PARP	BRCA mutation	✓
Osimertinib	EGFR	EGFR T790M mutation	✗
Palbociclib	CDK4, CDK6	HR+, HER2-	✓
Panitumumab	EGFR	KRAS wild type	✓
Pembrolizumab	PD-1	PD-L1+	✓
Pertuzumab	HER2	HER2+	✗
Ribociclib	CDK4, CDK6	HR+, HER2-	✓

Figure A.20 Assessment of driver targeting drugs for TCGA-AN-A0XN. This table contains driver-targeting drugs, i.e. those drugs that require the presence or absence of a specific mutation or other genomic alteration. The listed drugs are not necessarily approved for breast cancer and hence might be considered as off-label treatment options.

A.8.5.2 Case Study II: TCGA-BH-A0DT

TCGA sample of a 41-year-old (presumably premenopausal) woman with stage II breast cancer of TNM stage T1/N1/M0. Both hormone receptor (ER and PR) are positive, HER2 is not amplified. The tumor sample was predicted to be of *luminal A* subtype by the PAM50 classifier [727].

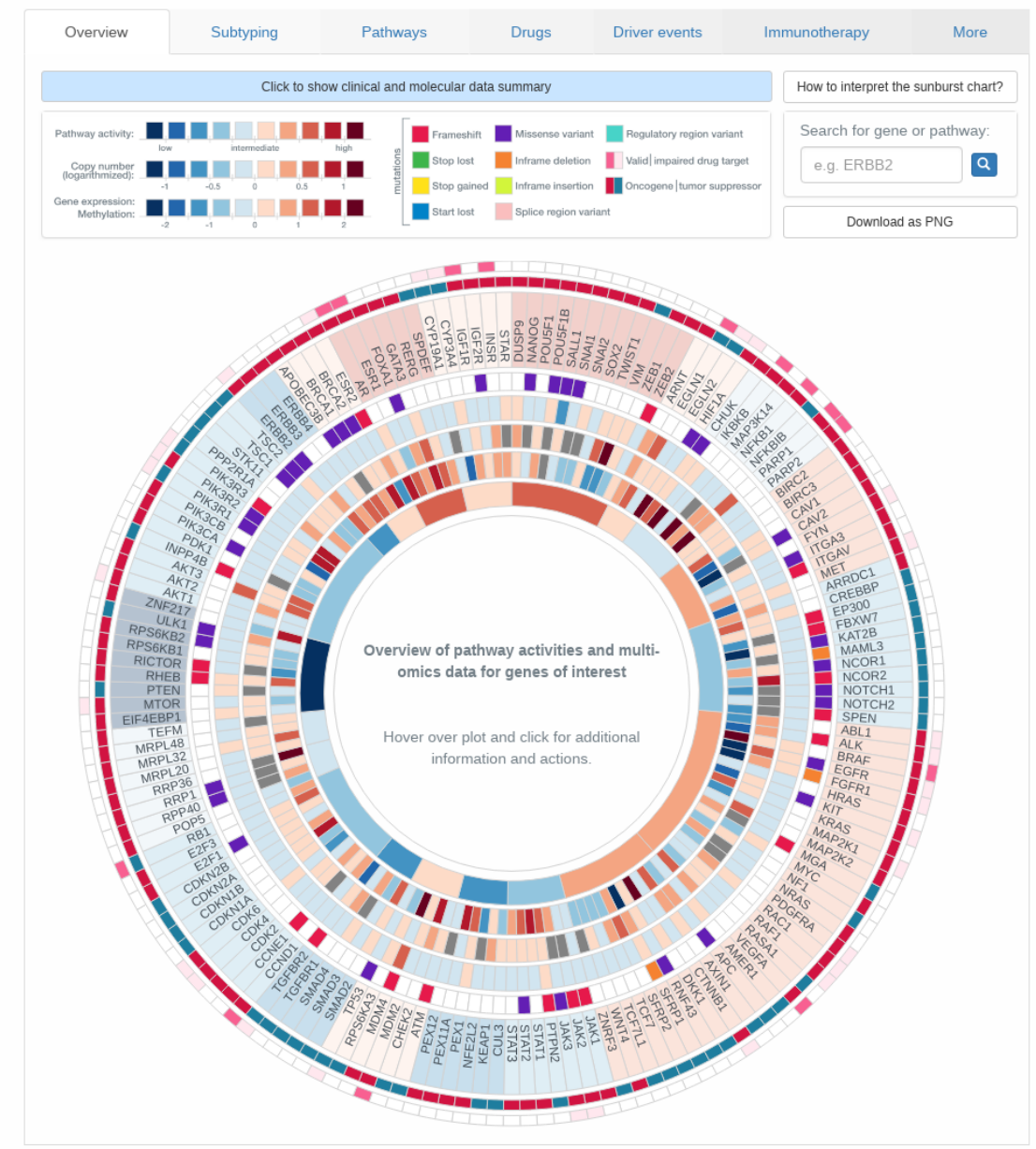


Figure A.21 Sunburst chart overview for TCGA-BH-A0DT. Breast cancer-relevant driver genes and pathways are displayed in a circular manner. Genes are grouped according to the pathways they are most characteristic for. The plot is organized in rings, where the innermost ring displays pathway activities, the second 'inner' ring corresponds to gene expression. Depending on the data provided by the user, information on methylation scores, copy number alterations, and mutations is shown in the third, fourth, and fifth ring, respectively. Gene names are displayed in the next ring. The second most outer ring indicates whether the gene acts as an oncogene or tumor suppressor gene (TSG) for activating the corresponding pathway. The outermost ring contains indicators on whether or not the gene is a known drug target.

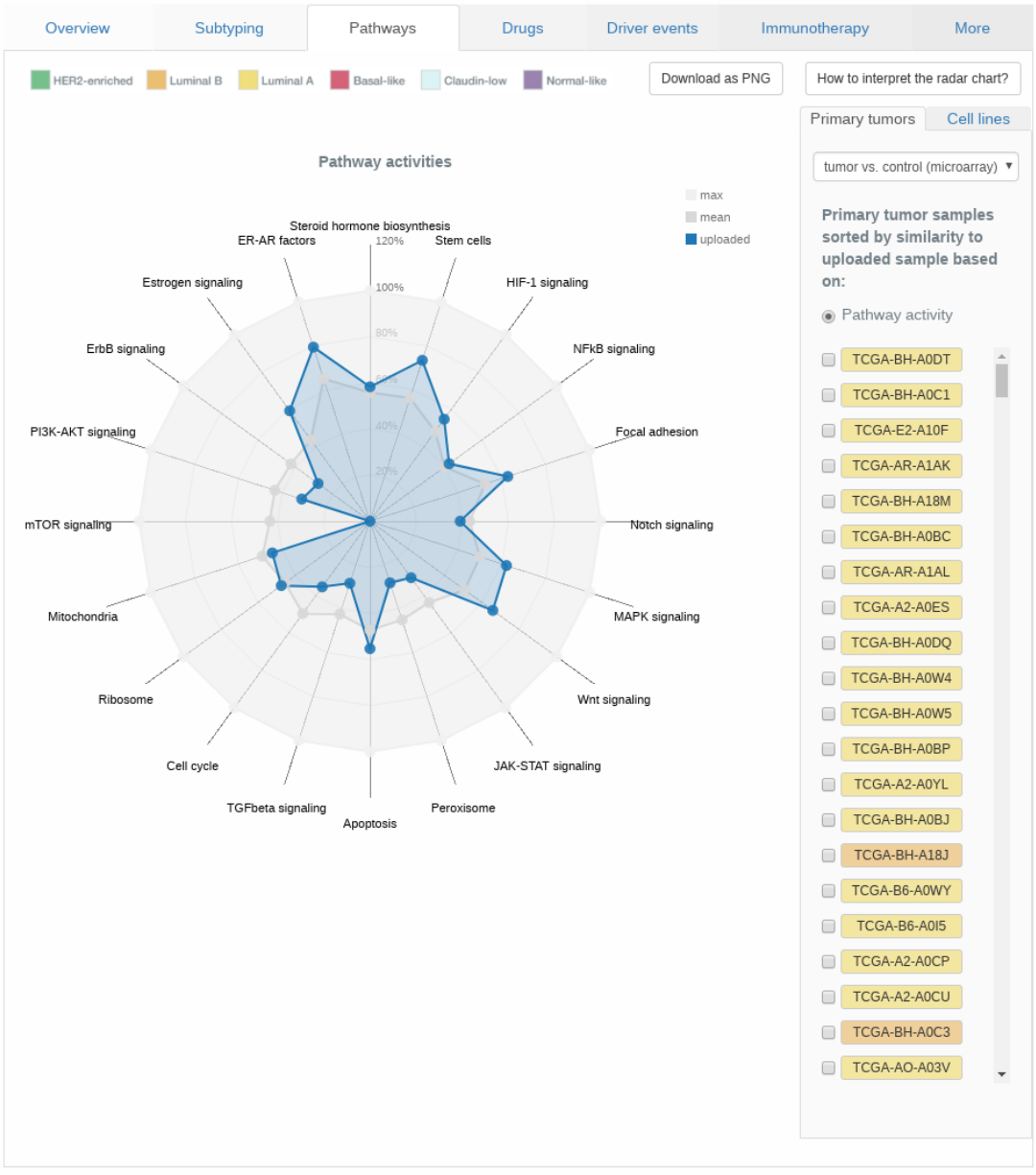


Figure A.22 Radar chart of pathway activities for TCGA-BH-A0DT. The pathway activities of a set of 20 core breast cancer pathways for the user-provided tumor sample colored in blue. Reference samples from TCGA as well as breast cancer cell lines can be added to the visualization interactively. The molecular subtype of the respective reference samples is color-coded: basal-like - red, claudin-low - light blue, HER2-enriched - green, luminal A - yellow, luminal B – orange, normal-like - purple.

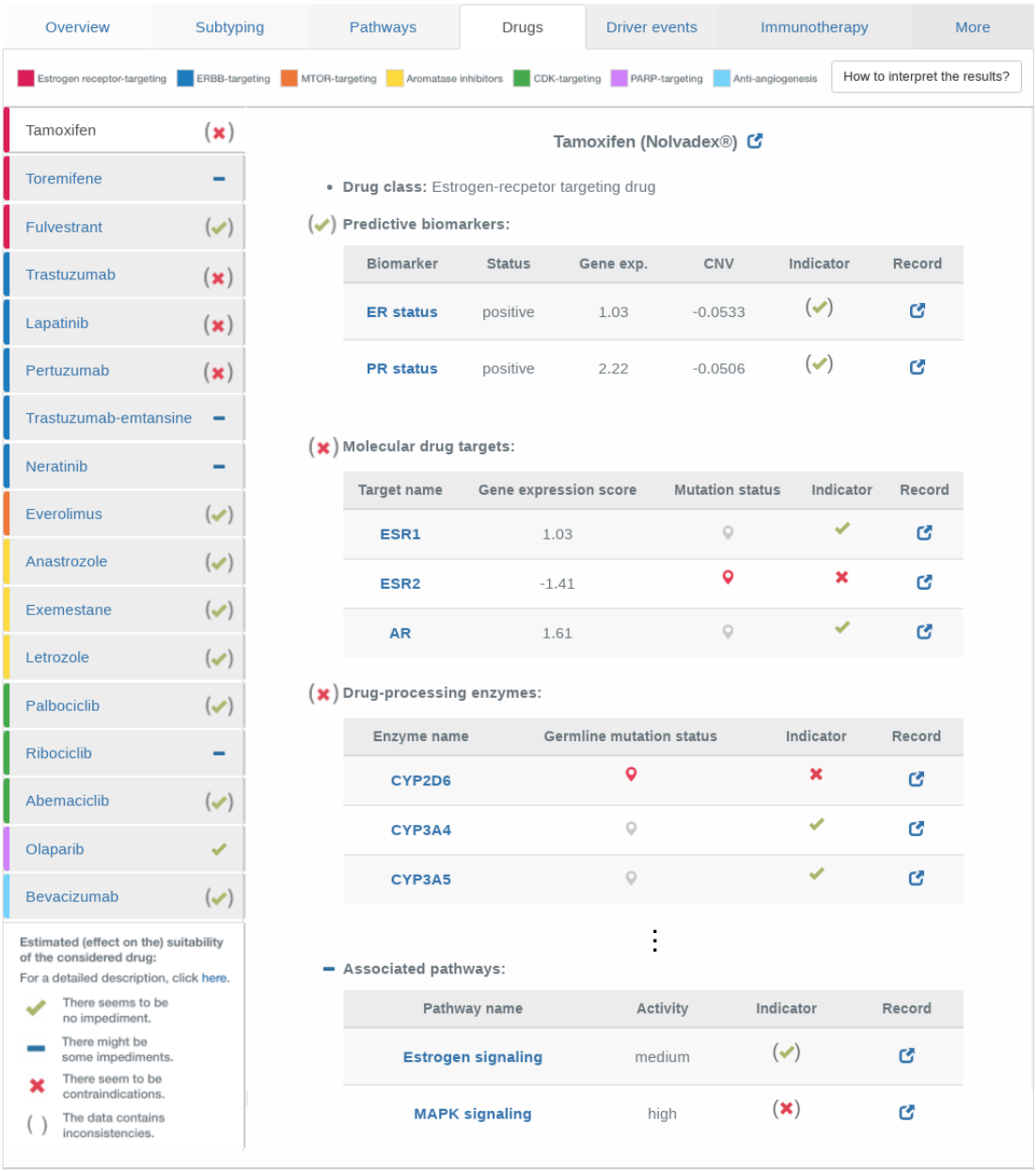


Figure A.23 **Assessment of tamoxifen for TCGA-BH-A0DT**. For a set of 17 standard-of-care breast cancer drugs (left panel), various factors increasing or decreasing the efficacy of a drug are assessed. Clinical, genetic, and molecular characteristics are listed with an indicator sign on whether they might decrease efficacy or even cause resistance to the treatment with the drug under consideration. All genes and pathways are linked to third-party resources where additional details can be found. Each entry also contains the link to a record or publication that describes the role of the corresponding gene with respect to the drug of interest. Clicking on the indicator symbol in the *Germline mutation status* column for CYP2D6 will open a window with additional details, see **Figure A.24**.

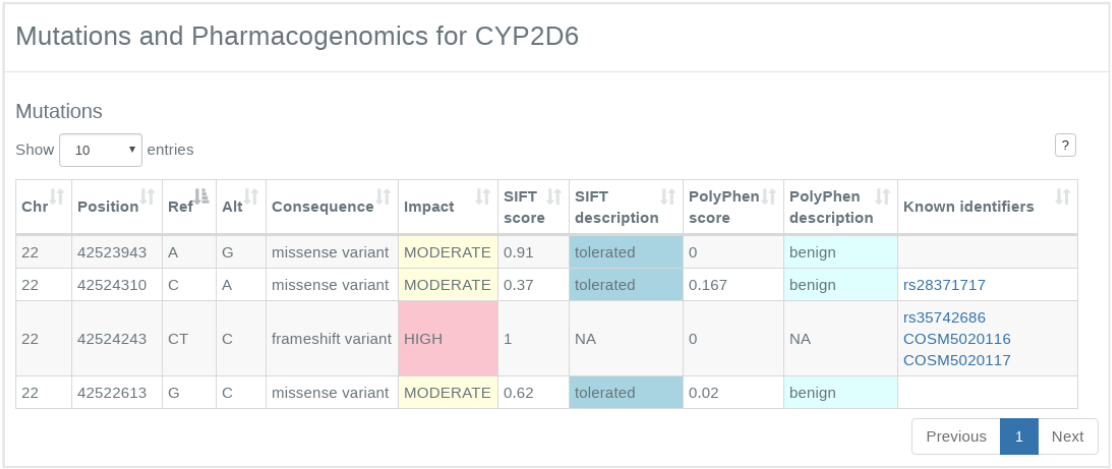


Figure A.24 **Detailed view on mutations in gene CYP2D6**. Clicking on the indicator symbol of a mutation opens a modal with additional details on the specific mutations contained in the gene of interest, as well as an estimation of the mutations severities based on VEP Impact, SIFT, and PolyPhen.

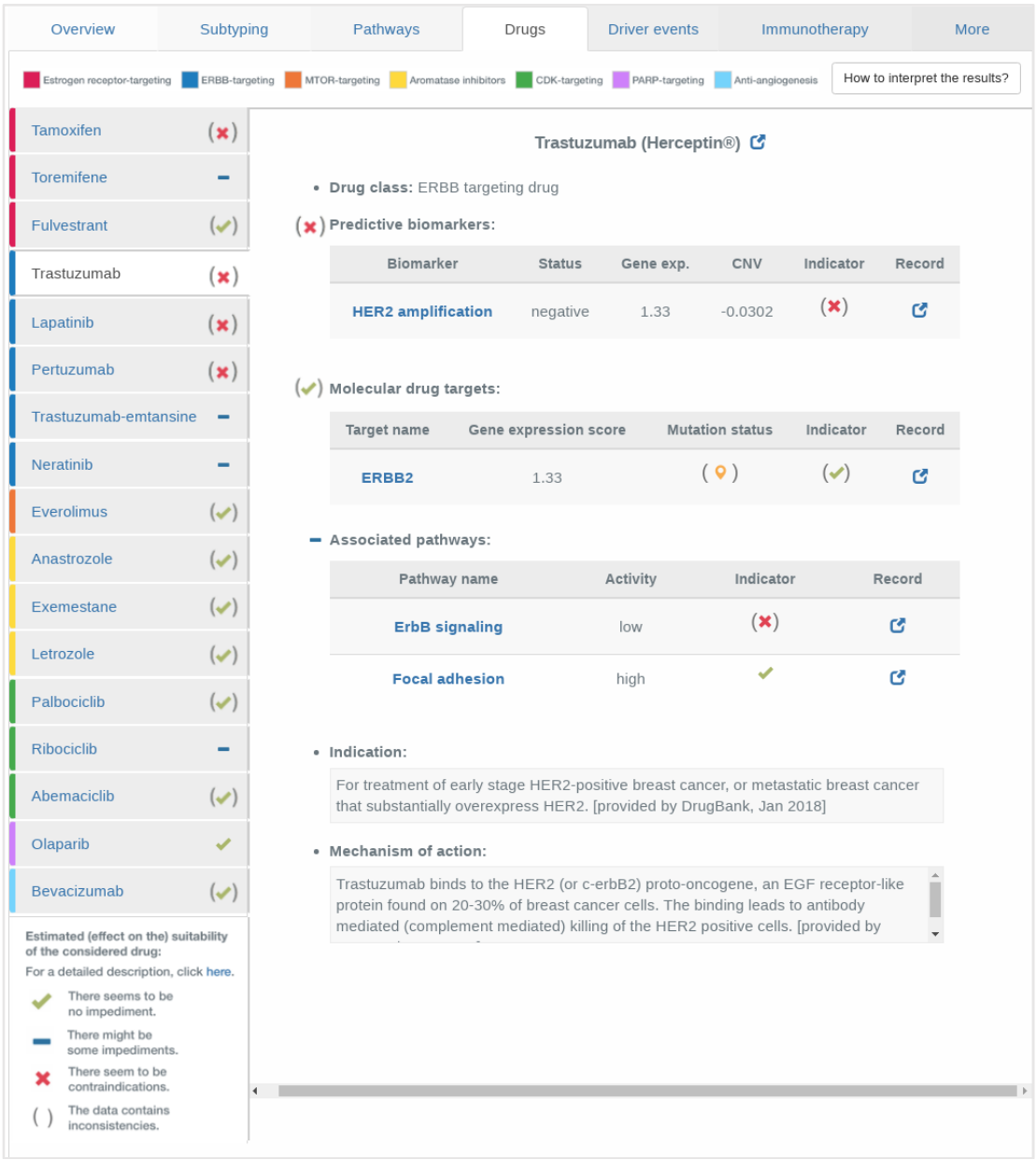


Figure A.25 Assessment of trastuzumab for TCGA-BH-A0DT. For a set of 17 standard-of-care breast cancer drugs (left panel), various factors increasing or decreasing the efficacy of a drug are assessed. Clinical, genetic, and molecular characteristics are listed with an indicator sign on whether they might decrease efficacy or even cause resistance to the treatment with the drug under consideration. All genes and pathways are linked to third-party resources where additional details can be found. Each entry also contains the link to a record or publication that describes the role of the corresponding gene with respect to the drug of interest.

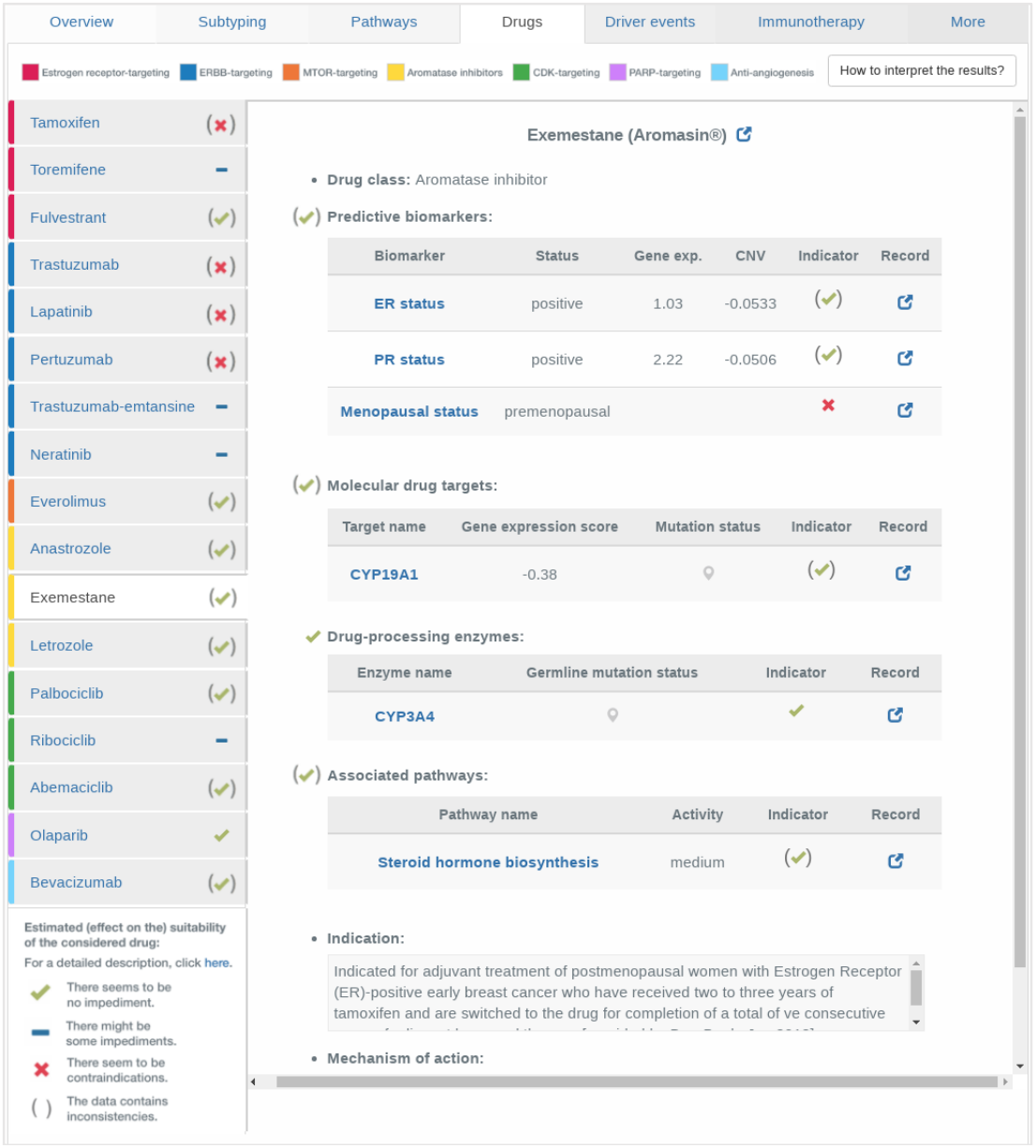


Figure A.26 Assessment of aromatase inhibitor exemestane for TCGA-BH-A0DT. For a set of 17 standard-of-care breast cancer drugs (left panel), various factors increasing or decreasing the efficacy of a drug are assessed. Clinical, genetic, and molecular characteristics are listed with an indicator sign on whether they might decrease efficacy or even cause resistance to the treatment with the drug under consideration. All genes and pathways are linked to third-party resources where additional details can be found. Each entry also contains the link to a record or publication that describes the role of the corresponding gene with respect to the drug of interest.

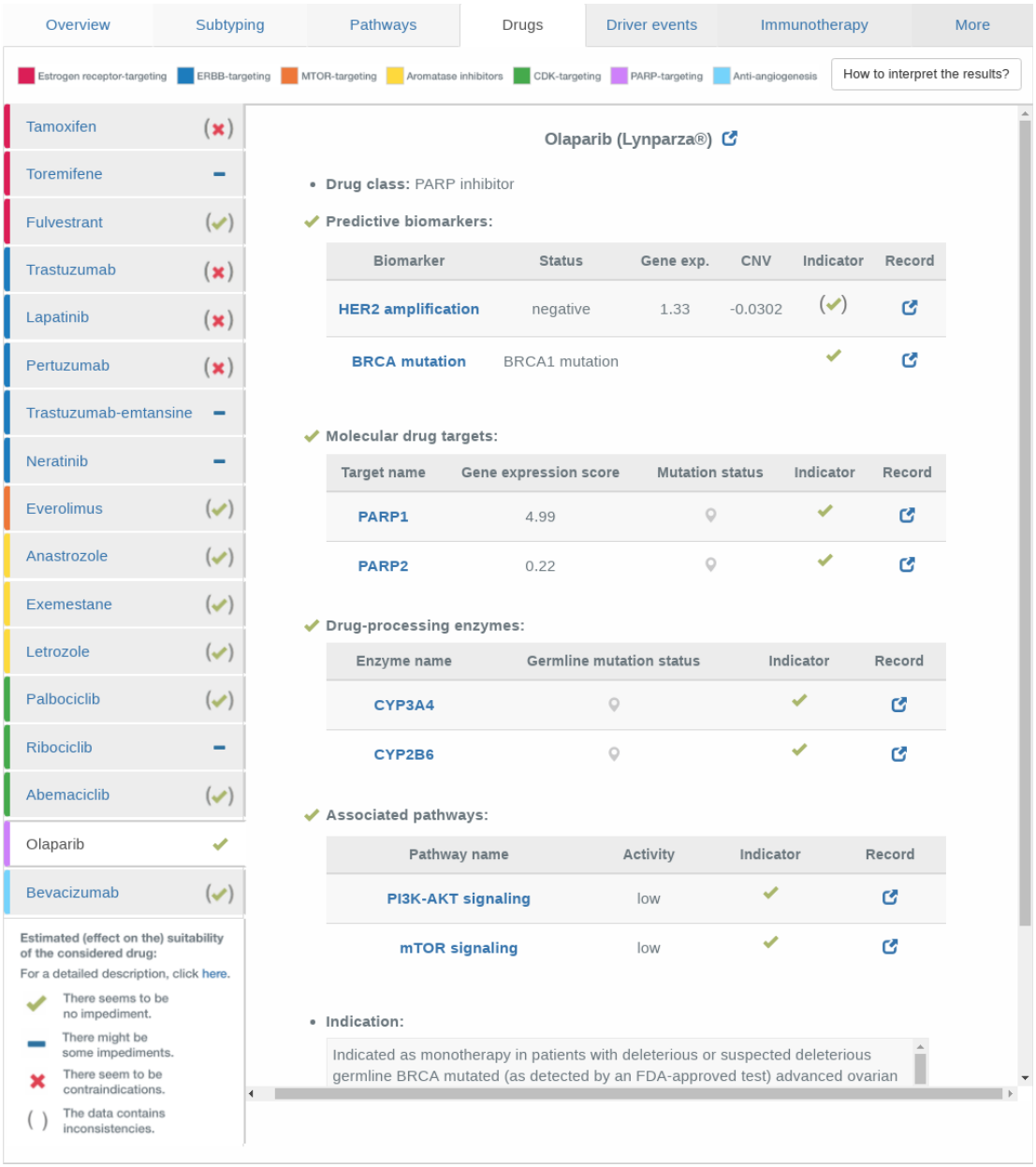


Figure A.27 Assessment of olaparib for TCGA-BH-A0DT. For a set of 17 standard-of-care breast cancer drugs (left panel), various factors increasing or decreasing the efficacy of a drug are assessed. Clinical, genetic, and molecular characteristics are listed with an indicator sign on whether they might decrease efficacy or even cause resistance to the treatment with the drug under consideration. All genes and pathways are linked to third-party resources where additional details can be found. Each entry also contains the link to a record or publication that describes the role of the corresponding gene with respect to the drug of interest.

A.8.5.3 Case Study III: TCGA-A2-A0T2

TCGA sample of a 66-year-old (presumably postmenopausal) woman with stage IV breast cancer of TNM stage T3/N3/M1. The sample is triple-negative (i.e., ER and PR negative, HER2 not amplified) and was predicted to be of *basal-like* subtype by the PAM50 classifier [727].

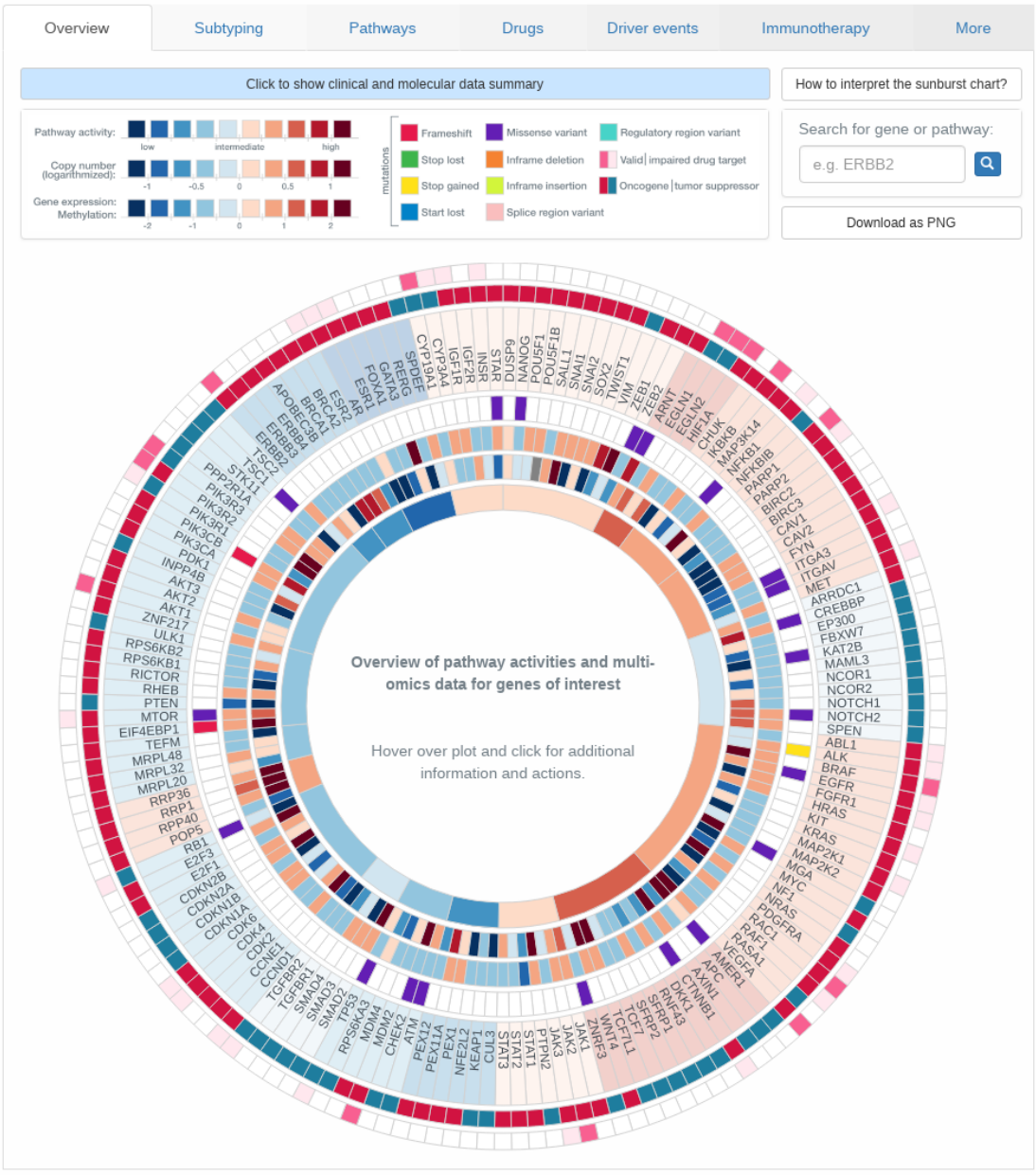


Figure A.28 Sunburst chart overview for TCGA-A2-A0T2. Breast cancer-relevant driver genes and pathways are displayed in a circular manner. Genes are grouped according to the pathways they are most characteristic for. The plot is organized in rings, where the innermost ring displays pathway activities, the second 'inner' ring corresponds to gene expression. Depending on the data provided by the user, information on copy number alterations and mutations is shown in the third and fourth ring, respectively. Gene names are displayed in the next ring. The second most outer ring indicates whether the gene acts as an oncogene or tumor suppressor gene (TSG) for activating the corresponding pathway. The outermost ring contains indicators on whether or not the gene is a known drug target.

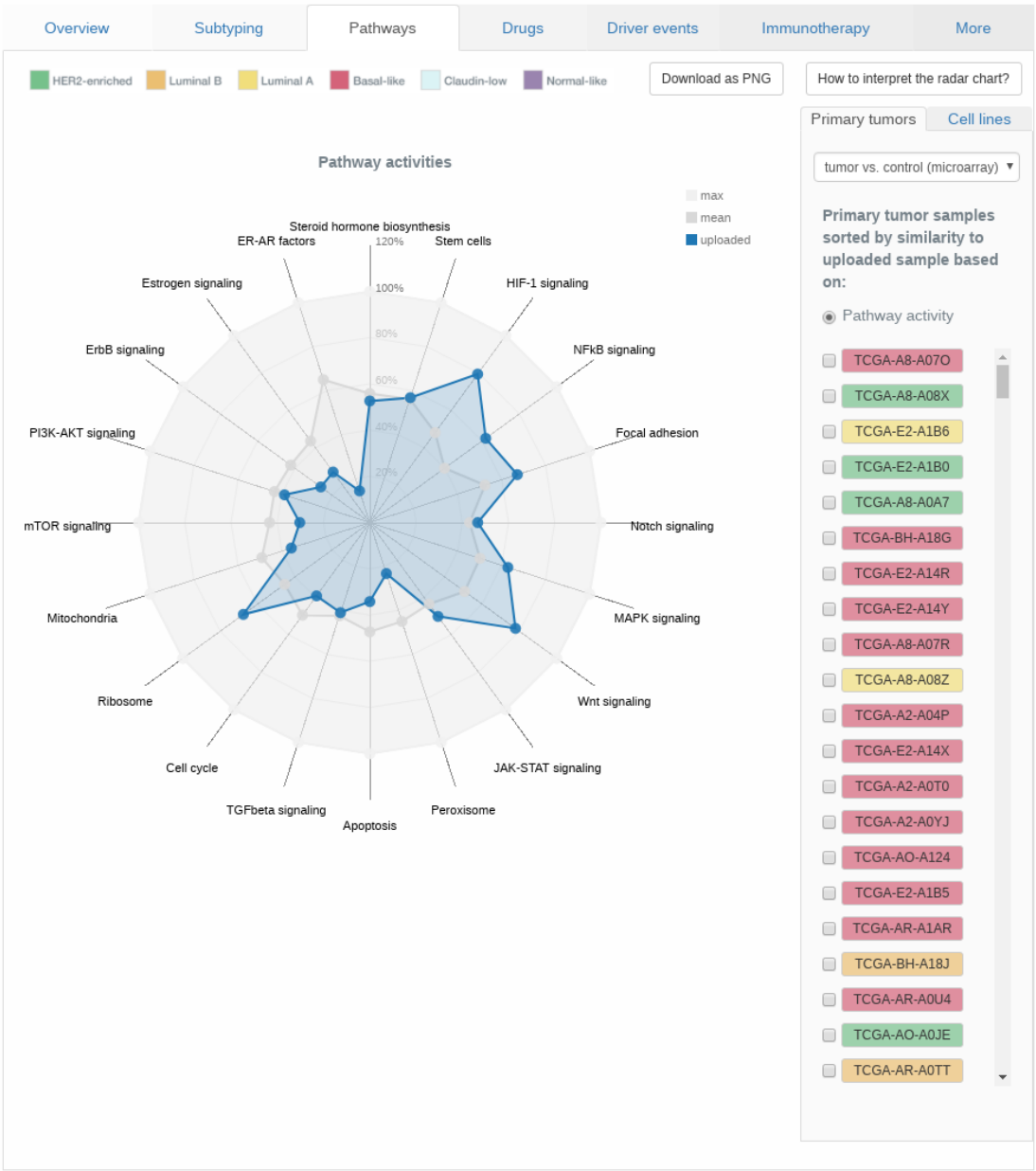


Figure A.29 Radar chart of pathway activities for TCGA-A2-A0T2. The pathway activities of a set of 20 core breast cancer pathways for the user-provided tumor sample colored in blue. Reference samples from TCGA as well as breast cancer cell lines can be added to the visualization interactively. The molecular subtype of the respective reference samples is color-coded: basal-like - red, claudin-low - light blue, HER2-enriched - green, luminal A - yellow, luminal B – orange, normal-like - purple.

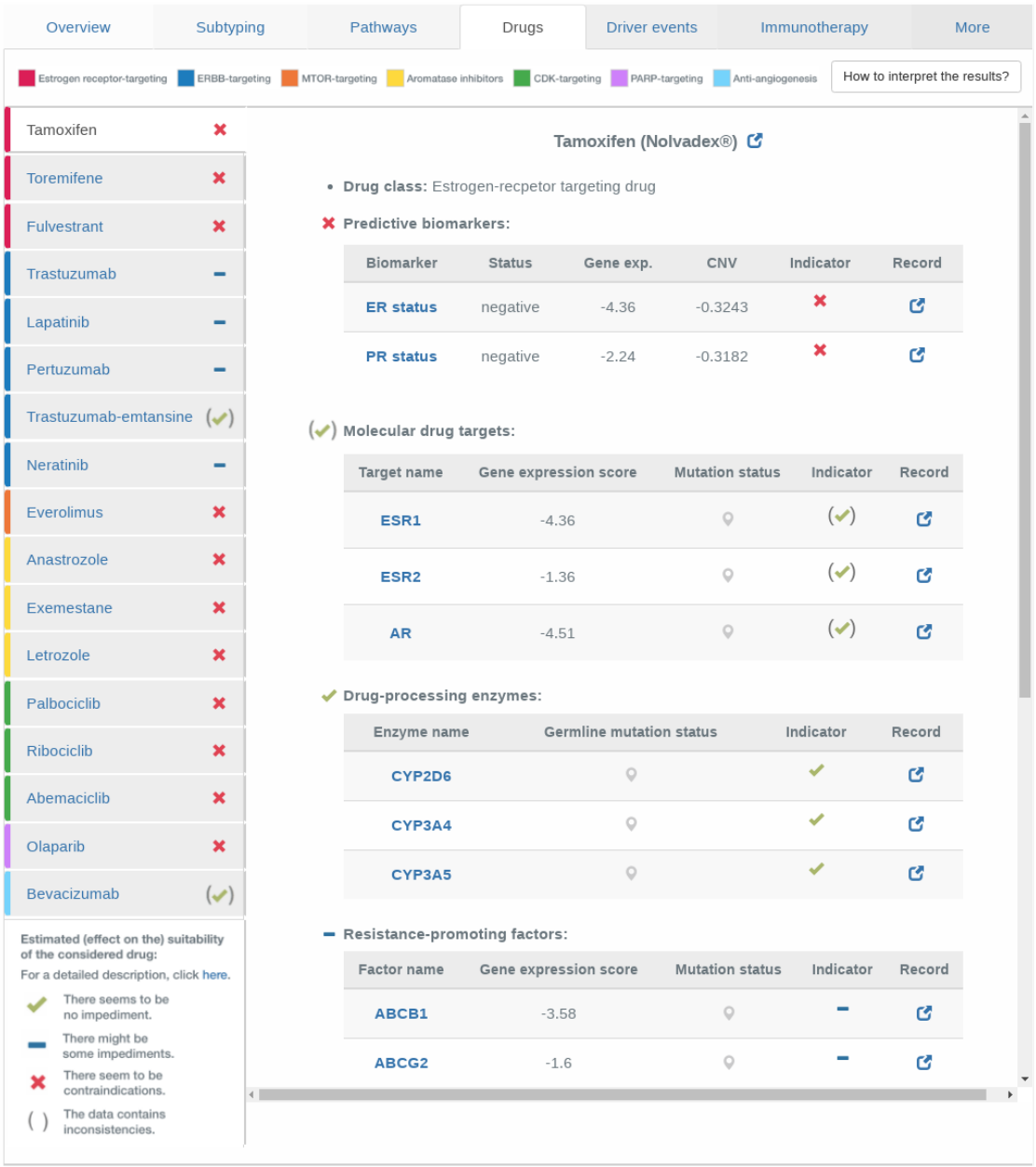


Figure A.30 Drug assessment for TCGA-A2-A0T2. For a set of 17 standard-of-care breast cancer drugs (left panel), various factors increasing or decreasing the efficacy of a drug are assessed. Clinical, genetic, and molecular characteristics are listed with an indicator sign on whether they might decrease efficacy or even cause resistance to the treatment with the drug under consideration. All genes and pathways are linked to third-party resources where additional details can be found. Each entry also contains the link to a record or publication that describes the role of the corresponding gene with respect to the drug of interest.

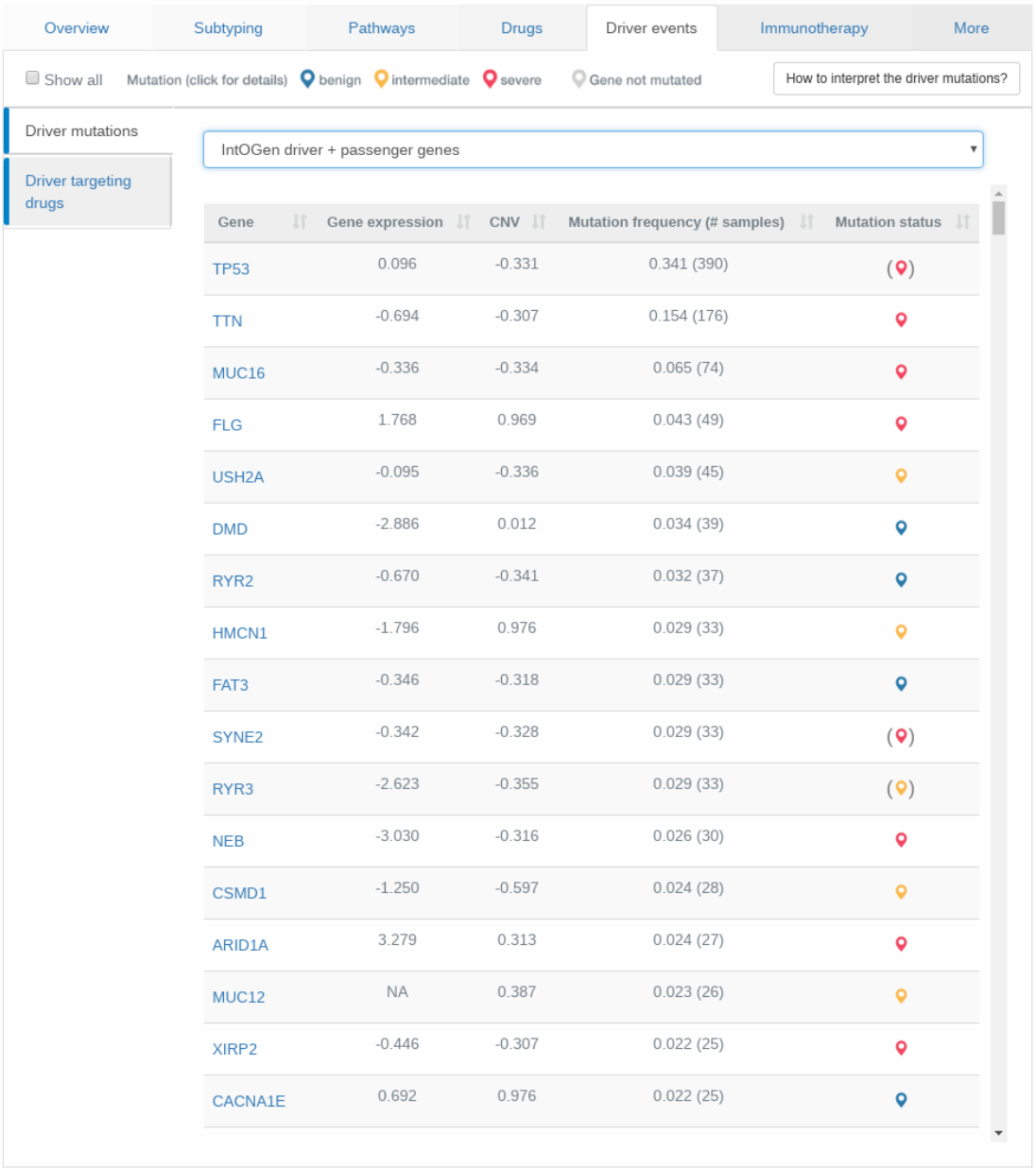


Figure A.31 Driver (and passenger) mutations in TCGA-A2-A0T2. The table contains genes commonly mutated in breast cancer samples that are also mutated in the sample under consideration. The mutations are sorted by decreasing frequency. The color-code in the *Mutation status* column indicates the severity of the contained mutations. Clicking on the respective symbol will open a modal with additional details on the contained mutations and their putative effect on protein functionality.

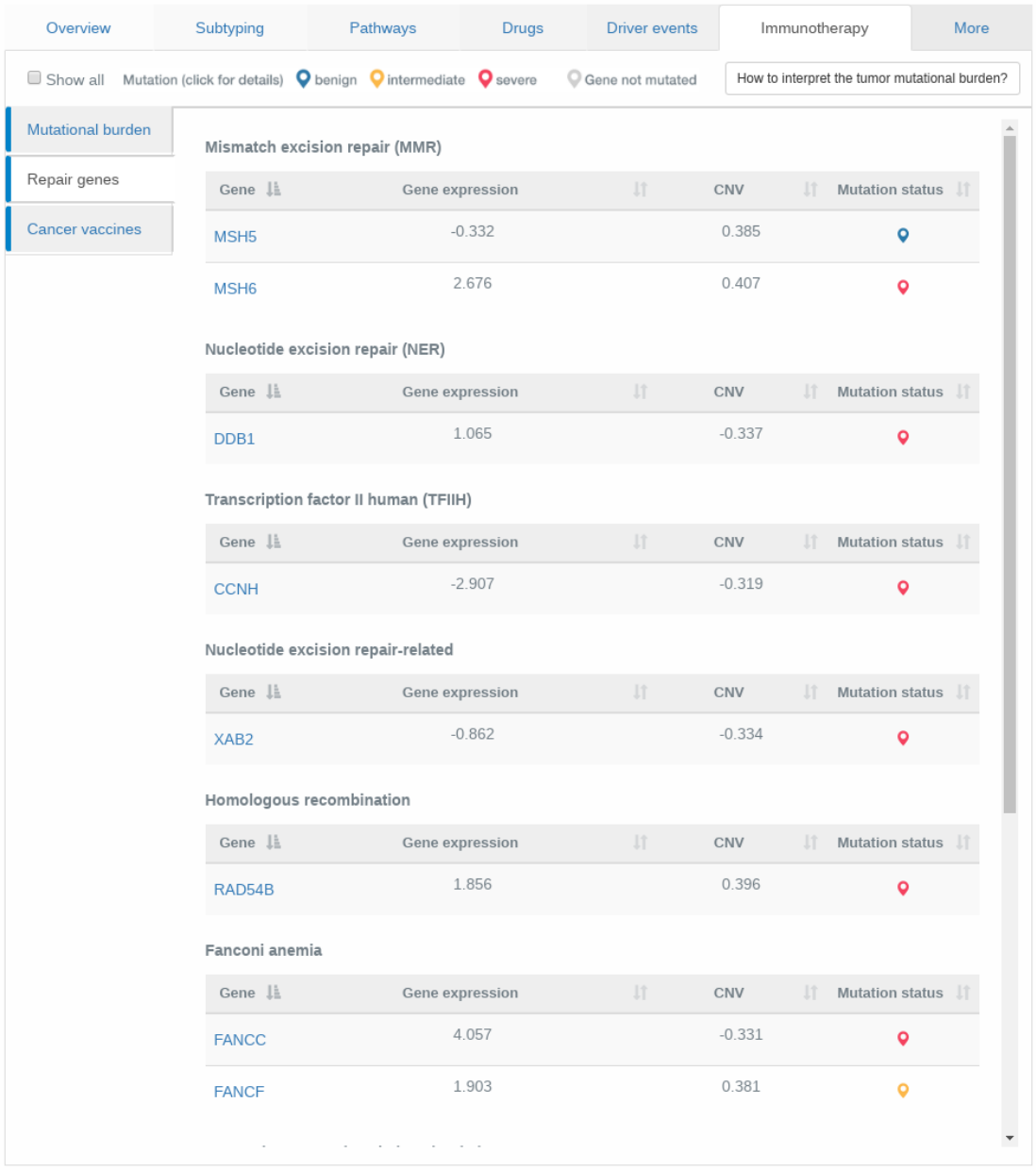
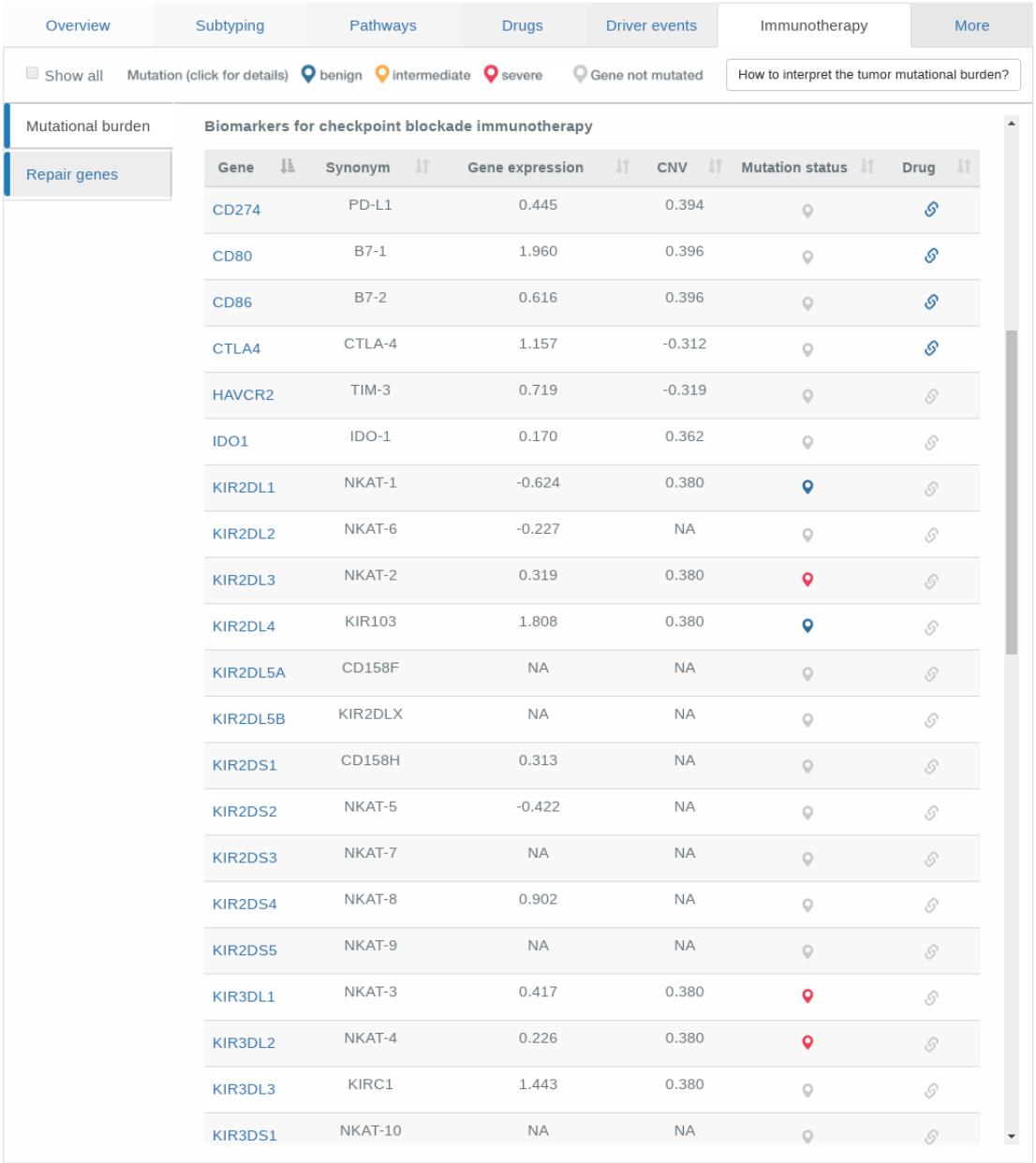


Figure A.32 Impaired repair genes in TCGA-A2-A0T2. The table contains genes involved in a variety of repair processes that are impaired (i.e., mutated) in the sample under investigation. The color-code in the *Mutation status* column indicates the severity of the contained mutations. Clicking on the respective symbol will open a modal with additional details on the contained mutations and their putative effect on protein functionality.



Gene	Synonym	Gene expression	CNV	Mutation status	Drug
CD274	PD-L1	0.445	0.394		
CD80	B7-1	1.960	0.396		
CD86	B7-2	0.616	0.396		
CTLA4	CTLA-4	1.157	-0.312		
HAVCR2	TIM-3	0.719	-0.319		
IDO1	IDO-1	0.170	0.362		
KIR2DL1	NKAT-1	-0.624	0.380		
KIR2DL2	NKAT-6	-0.227	NA		
KIR2DL3	NKAT-2	0.319	0.380		
KIR2DL4	KIR103	1.808	0.380		
KIR2DL5A	CD158F	NA	NA		
KIR2DL5B	KIR2DLX	NA	NA		
KIR2DS1	CD158H	0.313	NA		
KIR2DS2	NKAT-5	-0.422	NA		
KIR2DS3	NKAT-7	NA	NA		
KIR2DS4	NKAT-8	0.902	NA		
KIR2DS5	NKAT-9	NA	NA		
KIR3DL1	NKAT-3	0.417	0.380		
KIR3DL2	NKAT-4	0.226	0.380		
KIR3DL3	KIRC1	1.443	0.380		
KIR3DS1	NKAT-10	NA	NA		

Figure A.33 Biomarkers for checkpoint inhibition in TCGA-A2-A0T2. The table contains biomarkers for checkpoint blockade immunotherapy. In cases the listed genes are the molecular targets of immunotherapeutic drugs, the indicator mark in the *Drug* column is colored in blue. Clicking on this mark yields additional information on the targeting drugs.

Neoepitope	A*02:01	A*24:02	B*15:17	B*40:01	C*07:01	Antigen ID
AVWALCYGY	0.17	0.026	0.46	0.076	0.066	ZFP42
FKTTRIIFY	0.084	0.03	0.061	0.061	0.44	AGO2
FLLDMVYRS	0.771	0.05	0.031	0.042	0.074	DOPEY2
FSFGPQPYY	0.104	0.03	0.931	0.062	0.436	ROR1
FSPYNGGAL	0.075	0.059	0.519	0.137	0.16	ABTB2
GELINNTVL	0.057	0.027	0.054	0.841	0.052	GSDMC
GTSPSLIFL	0.437	0.058	0.514	0.198	0.239	SLC13A4
IGTPPSLIF	0.048	0.247	0.533	0.067	0.067	SLC13A4
IGTPTSLIF	0.06	0.301	0.535	0.077	0.087	SLC13A4
LAEVLAFLL	0.197	0.098	0.462	0.146	0.076	DOPEY2
LAFLLDMVY	0.118	0.032	0.741	0.076	0.096	DOPEY2
LAQKAIKQW	0.035	0.091	0.677	0.046	0.054	TRIM24
LLAEVLAFL	0.848	0.144	0.153	0.114	0.139	DOPEY2
MAFLAQKAI	0.167	0.052	0.691	0.089	0.254	TRIM24
MVFVAGQGV	0.426	0.035	0.674	0.125	0.307	SH3BP5L
QAACPPAIF	0.051	0.153	0.519	0.101	0.021	FANCC
RFKTTRIIF	0.048	0.44	0.101	0.096	0.087	AGO2
RVILAKRLY	0.055	0.035	0.649	0.074	0.265	AXDND1
STRFKTTRI	0.066	0.081	0.634	0.062	0.149	AGO2
TIIGTPPSL	0.433	0.113	0.248	0.11	0.067	SLC13A4
TIIGTSPSL	0.451	0.12	0.356	0.128	0.093	SLC13A4
TRFKTTRII	0.044	0.094	0.047	0.08	0.659	AGO2
VLAFLLDMV	0.674	0.083	0.042	0.084	0.039	DOPEY2
VSAVWALCY	0.079	0.077	0.87	0.061	0.17	ZFP42
WEIFSFGPQ	0.049	0.014	0.021	0.583	0.009	ROR1
WTGWVCCVF	0.144	0.257	0.49	0.114	0.041	CENPL
YCTGPCHTF	0.074	0.369	0.571	0.151	0.139	PPP2R3A
YSGGEKPYL	0.192	0.078	0.479	0.078	0.29	COPB2
YSRIPKQSI	0.052	0.062	0.833	0.061	0.196	SVIL
YVVTEAGEL	0.223	0.05	0.441	0.132	0.158	GSDMC
YWQGNLDRF	0.061	0.655	0.111	0.093	0.09	POLR3C

Table A.15 **Neoepitope prediction for sample `TCGA-A2-A0T2`.** The table contains all neoepitopes of peptide length nine that were predicted by NetMHC to bind to at least one of the sample's predicted HLAs (predicted by OptiType). Within ClinOmicsTrail[bc], we selected the option *Consider only significantly upregulated proteins* in the *Cancer vaccines* tab of the *Immunotherapy* view and performed neoepitope prediction for peptides of length 9 using NetMHC. For each HLA, their respective binding affinities are listed in columns 2-6. The last column contains the gene symbol of the antigen-providing gene. For NetMHC, binding affinity scores can be interpreted as transformed IC50 scores. Red cells indicate no binding, yellow cells stand for weak binders, and green cells highlight strong binders.