Prediction and visualization of the carcinogenic potential of chemicals with short-term *omics* assays

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

MSc. Bioinform. Michael Römer

aus Räckelwitz

Tübingen

2017

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	24.05.2017
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Andreas Zell
2. Berichterstatter:	Prof. Dr. Michael Schwarz

Abstract

Drug candidates that induce or promote cancer formation must be identified and eliminated during the preclinical phase of drug development to minimize the risk of adverse, carcinogenic effects in patients. Genotoxic carcinogens can be identified with short-term assays. In contrast, the lifetime rodent cancer bioassay that is used to identify nongenotoxic carcinogenic substances, requires a high number of test animals and takes up to five years for completion. In addition, the lifetime rodent cancer bioassay does not provide sufficient data to evaluate the human risk if carcinogenic effects are observed in rodents. This can result in discontinuation of the development of the drug candidate or a black label warning on the drug packaging. The application of high-throughput *omics* methods such as transcriptomics or proteomics in toxicological studies is a promising approach for the development of short-term alternatives to the lifetime rodent cancer bioassay. However, these *omics* methods are difficult to use for life sciences researchers and few specialized visualization tools exist for toxicogenomics data. Furthermore, most existing studies used only a single *omics* platform to determine the molecular effects of carcinogens.

This thesis introduces new approaches that integrate multiple *omics* platforms for the identification of nongenotoxic carcinogens and presents analysis and visualization tools that were specifically developed for toxicogenomics data. We performed a series of experiments to demonstrate that our multi-*omics* approach improves the prediction performance compared to single-*omics* approaches. To facilitate the access to our analysis and visualization tools, we implemented two web platforms, the ZBIT Bioinformatics Toolbox and MARCARviz. These web platforms enable toxicologists to gain new insights into the mechanisms of nongenotoxic tumor promotion. Furthermore, we demonstrated that our multi-*omics* approach can provide the basis of new short-term alternatives to the lifetime rodent cancer bioassay.

Kurzfassung

Arzneimittelkandidaten die die Entstehung und das Wachstum von Tumoren begünstigen, müssen in der präklinischen Phase der Medikamentenentwicklung identifiziert und aus der weiteren Entwicklung ausgeschlossen werden, um das Risiko von gefährlichen, tumorfördernden Nebenwirkungen für Patienten zu minimieren. Während genotoxische Substanzen mit Schnelltests identifiziert werden können, dauert das aktuelle Standardprüfverfahren zur Erkennung von nicht-genotoxischen, karzinogenen Substanzen bis zu fünf Jahre und benötigt eine große Anzahl an Versuchstieren. Außerdem können aus dem Ergebnis keine Hinweise auf den Mechanismus gezogen werden wenn bei der Prüfung Tumore gefunden werden, was zur Einstellung der Entwicklung des Arzneimittelkandidaten oder zu einer Black-Box-Warnung auf der Verpackung führen kann. Die Anwendung von modernen Hochdurchsatz-Technologien in toxikologische Studien, Toxikogenomik genannt, ist ein vielversprechender Ansatz zur Entwicklung von Prüfverfahren, die weniger Zeit und Versuchstiere benötigen. Allerdings sind die Methoden aus der Toxikogenomik für Toxikologen oft schwierig anzuwenden. Außerdem berücksichtigten die meisten existierenden Studien nur Daten einer einzelnen *omics*-Technologie und es existieren nur wenige spezialisierte Visualisierungswerkzeuge für toxikogenomische Daten.

Diese Arbeit stellt neue Analyse- und Visualisierungswerkzeuge vor, die spezifisch für toxikogenomische Studien entwickelt wurden, sowie integrative Ansätze, die es ermöglichen Daten von mehreren *omics*-Plattformen zu berücksichtigen, um die Identifikation von nicht-genotoxischen Karzinogenen zu verbessern. Wir beschreiben eine Reihe von Experimenten mit einem neuen Toxikogenomikdatensatz, um zu demonstrieren, dass unsere integrativen Ansätze die Vorhersage der Karzinogenität von Substanzen verbessern. Die Weiterentwicklung der von uns beschriebenen integrativen Verfahren bietet möglicherweise Alternativen zu dem aktuell verwendeten, zeitaufwändigen Verfahren zur Feststellung der Karzinogenität. Außerdem beschreiben wir neue Webplattformen zur Analyse und Visualisierung von Expressionsdaten aus der Toxikogenomik, die wir entwickelt haben, um Toxikologen den Zugang zu bioinformatischen Werkzeugen zu vereinfachen. Mit diesen neuen Webplattformen können Toxikologen neue Erkenntnisse über die Wirkmechanismen der nicht-genotoxischen Krebsentstehung gewinnen.

Contents

Chapter 1

Introduction

The discovery of drugs for common and rare diseases has improved the prognosis of many diseases and greatly extended human life expectancy. The development and marketing of pharmaceutical drugs is a multi-billion dollar industry. Based on the data of 106 randomly selected new drugs, DiMasi *et al.* (2016) report that pharmaceutical companies invest over 10 years of research and $2.5 billion until a new drug is approved for marketing. This estimate includes the cost of drug candidates that are discontinued during the development or risk assessment process due to lack of efficacy or severe adverse side effects. During the premarketing risk assessment, drug candidates are screened for adverse side effects in a process that is divided into several phases. The preclinical phase involves *in silico*, *in vitro*, and animal testing. Three subsequent clinical phases involve increasing numbers of patients and healthy participants to evaluate the efficacy and side effects of drug candidates in humans. Based on industry reports, DiMasi *et al.* (2016) calculate a clinical success rate of 11.83%, which means that 88.17% of the compounds that pass the preclinical phase fail during one of the three clinical phases. When drug candidates pass clinical phase III, regulatory agencies approve them for marketing. Approved and marketed drugs enter clinical phase IV, the postmarketing surveillance. Physicians and regulatory agencies continue to monitor the effects of the drug in patients and aggregate new results from further animal experiments. Onakpoya *et al.* (2016) report that manufacturers and regulatory agencies withdrew 462 approved medicinal products due to adverse drug reactions between 1953 and 2013. Hepatotoxicity (81 cases) was the most common reason for withdrawals, and 114 of the 462 withdrawals were associated with patient deaths. 61 drugs were withdrawn due to carcinogenicity. For example, the antihistamine methapyrilene that was used in flu medicines was withdrawn after it was found to be a potent liver carcinogen in animal carcinogenicity studies (Lijinsky *et al.*, 1980).

Animal studies are an integral part of toxicological research and the drug discovery process. Mechanistic studies in animals establish biological pathways and determine druggable targets for diseases. These druggable targets are the basis for modern drug development. High-throughput chemical screenings identify small molecules (called leads) that show a binding affinity for the druggable target. Molecular optimization increases the binding affinities through modifications of the structure and composition of the lead.

Further *in vitro* and *in vivo* animal studies establish the efficacy, toxicity, and pharmacokinetics of successful leads and act as a filter for the initial drug candidates. Estimates by Bolten and DeGregorio (2002) indicate that only 1 in 1,000 compounds that pass the early stages of drug discovery enter the clinical phase. Due to the high number of compounds that fail before entering the clinical phase, pharmaceutical companies perform the most expensive and time-consuming animal tests late in the development process, long after clinical trials have started (Kramer *et al.*, 2004). One of these tests is the lifetime rodent cancer bioassay (LRB). It requires more than 800 mice and rats, substantial amounts of the drug candidate, and the histopathological examination of more than 40 tissues (Waters *et al.*, 2010). Regulatory agencies demand the LRB for drug candidates that are to be administered to patients chronically for more than six months. Tumor findings in the LRB may cause significant delays in the approval of drug candidates or even result in the withdrawal of approved drugs, as was the case for methapyrilene (Lijinsky *et al.*, 1980).

Because most compounds that cause DNA damage are potent carcinogens, pharmaceutical companies eliminate genotoxic compounds early in the development. However, some carcinogens, such as methapyrilene, initiate or promote tumors but do not cause DNA damage. These compounds are called nongenotoxic carcinogens (NGCs). In consequence, the *in vitro* genotoxicity assays that are performed in early stages of drug development fail to detect their carcinogenic potential. NGCs are the main reason for tumor findings in the LRB. The relevance of these positive findings to humans is a controversial issue, and the LRB does not provide mechanistic information to determine the mode of action (MOA) through which a compound induces cancer (Ames and Gold, 1990). For these reasons, as well as the high cost in time and money and more recent political initiatives to reduce the number of animal tests in risk assessment, pharmaceutical companies and regulatory agencies are working towards alternative approaches for assessing the carcinogenic potential of new drugs. The innovative medicines initiative (IMI) MARCAR project, which encompasses several European universities and pharmaceutical companies, is a combined effort to identify molecular biomarkers and tumor classifications for nongenotoxic carcinogenesis. The MARCAR Consortium (2010) formulates four major objectives: (i) Find early biomarkers that can reliably identify compounds with potential for later cancer development, (ii) advance the scientific basis for the assessment of the carcinogenic potential of nongenotoxic drugs, (iii) identify the molecular responses to the exposure to NGCs to support the development of early biomarkers, and (iv) improve the efficiency of risk assessment in drug development by progressing the development of alternative research methods. The MARCAR consortium expects that early cancer biomarkers improve the safety of participants in clinical trials and reduces the need for animals in accordance with the proposed concept of reduction, refinement, and replacement of animal experimentation (MARCAR Consortium, 2010).

The inclusion of bioinformatics and high-throughput technologies is a central component of most proposed alternatives to the LRB. These so-called *omics* methods contribute a rich source of molecular data, which toxicologists use to perform mechanistic analy-

ses. Machine learning models extract biomarker signatures from this high-dimensional data and predict the carcinogenicity of compounds. The field that combines toxicological questions with *omics* methods has been coined toxicogenomics. The maturation of microarray technologies and next-generation sequencing simplified the collection of molecular data for many compounds. This led to the creation of two large standardized databases of the effects of NGCs on gene expression in rodent tissues: the Toxicogenomics Project-Genome Assisted Toxicity Evaluation System (TG-GATEs) by Uehara *et al.* (2010) and DrugMatrix by Ganter *et al.* (2006). Prompted by the need for alternative strategies for assessing the carcinogenic potential and the advances of *omics* technologies, the MARCAR project explores the application of multi-*omics* and epigenetic profiling in preclinical risk assessment. To this end, the MARCAR consortium profiled gene expression, protein abundance, microRNA (miRNA) expression, and DNA methylation in the same samples to generate an integrated model of the effects of NGCs on all levels of gene regulation (see, for example, Thomson *et al.* (2014) and Unterberger *et al.* (2014)). To efficiently analyze this data, new integrative bioinformatics methods and specialized visualization tools are necessary.

1.1 Contributions of this thesis

This thesis presents the research that we conducted in cooperation with members of the chair of Cognitive Systems and the MARCAR consortium. The presented research focuses on the visualization and predictive analysis of the toxicogenomics data generated during the MARCAR project. Due to the integrative approach of the MARCAR project and the high-dimensional nature of *omics* data, the analysis requires automated processing pipelines, efficient data storage infrastructure, robust statistics, and advanced machine learning methods. To facilitate data analysis and access to the data for the whole MARCAR consortium, we developed web platforms which provide visualizations and analysis tools for the MARCAR data. In addition, we developed new toxicogenomics approaches that were designed specifically for the integrative data collected by the MARCAR project. The following paragraphs shortly summarize the four main contributions of this thesis.

The ZBIT Bioinformatics Toolbox for computational biology

Due to the advance of high-throughput technologies in the life sciences, computational data analysis is an integral part of modern research. However, the computational resources and the technical knowledge of many wet lab researchers are limited, whereas the requirements in both regards are rising with each new generation of *omics* technologies. For example, the installation of academic bioinformatics software often depends on specific operating systems or third-party libraries. Also, many tools do not provide graphical user interfaces or detailed documentation. In addition, the processing *omics*

data can require large amounts of RAM and computational power, which are not available on standard desktop computers. For these reasons, we set up the ZBIT Bioinformatics Toolbox, which provides web-based access to a collection of bioinformatics software and pipelines. By porting bioinformatics software into the web, we enable researchers to use our tools on our computation cluster, such that they do not need to install any dependencies or worry about hardware restrictions. Currently, the software collection in the ZBIT Bioinformatics Toolbox encompasses tools for systems biology, expression data analysis, and transcription factor annotation. We use a customized version of the Galaxy framework to host the tools and distribute the tool execution among the nodes of our internal computation cluster. A workflow system allows the combination of tools into automated pipelines and the ZBIT Bioinformatics Toolbox stores all results and parameters in accordance with scientific needs for persistence and reproducibility. As a consequence, the user requires only a browser to access the tools, perform analysis, and view results in standardized formats. We provided extensive documentation for all tools, together with tutorials, use cases, and example data. The ZBIT Bioinformatics Toolbox is freely available at `https://webservices.cs.uni-tuebingen.de/` and was published in *PLoS ONE* in 2016 (Römer *et al.*, 2016b).

Similarity screening for characterization of drug candidates

ToxDBScan is a tool for evaluating the hepatocarcinogenic potential of drug candidates or other compounds in rodents. Users can upload gene expression signatures, which ToxDBScan uses to identify substances that induce similar changes. Based on the known properties and mechanisms of these similar substances, users can extrapolate information on hepatocarcinogenicity and potential modes of action. ToxDBScan uses a novel similarity screening method, which requires only the up- and downregulated genes as input. The ranking based on the new similarity scoring achieved a sensitivity of 88% and a predictive analysis correctly predicted the carcinogenicity of 15 external validation compounds. Furthermore, ToxDBScan visualizes the most similar expression patterns in heat maps and performs a pathway enrichment analysis for the gene expression signature to provide the user with additional mechanistic information. ToxDBScan is freely available from the ZBIT Bioinformatics Toolbox and was published in the *International Journal of Molecular Sciences* in 2014 (Römer *et al.*, 2014b).

Multi-omics approaches for prediction of nongenotoxic carcinogenicity

Traditionally, predictive toxicogenomics studies use mRNA microarrays to determine molecular signatures for the early assessment of the carcinogenic potential of new compounds. While some studies explore the use of other *omics* platforms such as miRNA microarrays or proteomics arrays, most existing models are based on data from a single *omics* platform. Here, we explored the integration of data from multiple *omics* platforms to build predictive models for compound carcinogenicity. To this end, mRNA,

miRNA, and protein expression profiles were collected from the livers of rats, which were exposed to NGCs, genotoxic carcinogens (GCs), and noncarcinogens (NCs). We developed new integrative feature representations which provide an abstraction from the molecular perspective of traditional toxicogenomics models to a systematic, pathway- and interaction-based perspective. These feature representations are calculated based on the observed expression values and incorporate existing knowledge that is available from molecular interaction databases and pathway databases, such as KEGG, BioCarta, or Reactome. We evaluated the performance of models for hepatocarcinogenicity prediction, which were built with five different machine learning methods and several combinations of the multi-*omics* features. With a repeated cross-validation procedure, we found that the integration of data from multiple *omics* platforms increases the prediction accuracy in hepatocarcinogenicity classification. We also demonstrated that the classification performance increases further when the proposed integrative feature representations are available for classifier training. In consequence, we were able to demonstrate that the early identification of NGCs can be improved by profiling and integrating data from multiple *omics* platforms. This study was published in *PLoS ONE* in 2014 (Römer *et al.*, 2014a).

Web-based interactive visualization of gene regulation data

The effective mining of high-dimensional toxicogenomics datasets is a non-trivial task that usually requires bioinformatics support to extract relevant mechanistic patterns and confirm toxicological hypotheses. MARCARviz is a web platform that enables biologists to quickly address the most common questions associated with the MARCAR microarray data, to identify relevant patterns in the data, and to generate or confirm mechanistic hypotheses about nongenotoxic effects leading to cancer formation. The major advantage of MARCARviz is that there is no software or advanced technical knowledge required to perform powerful analyses and generate visualizations of the MARCAR data. MARCARviz greatly facilitates the confirmation of published MARCAR results and generation of new insights from the collected data without the requirement for complex preprocessing steps. MARCARviz is publicly available from `https://tea.cs.uni-tuebingen.de/` and was presented at the German Conference for Bioinformatics 2016 (Römer *et al.*, 2016a).

1.2 Thesis structure

The following two chapters provide the biological and statistical background for the research that is presented in this thesis. The first half of Chapter 2 describes the levels of gene regulation that are relevant for nongenotoxic carcinogenicity and have been targeted by molecular profiling in the MARCAR project, followed by a description of the high-throughput technologies that were used to collect the data used in this thesis. The second half summarizes the current knowledge of the mechanisms of chemical-induced

carcinogenesis with particular focus on NGCs. Chapter 3 provides an overview of the preprocessing and statistical analysis of gene regulatory *omics* data. Chapter 3 also introduces the machine learning algorithms that have been used in the studies presented in Chapters 5 and 6 and the validation strategies that are used to assess the classification performance. The last section of Chapter 3 reviews the current state of predictive toxicogenomics, its anticipated impact on preclinical risk assessment, and the limitations of recent toxicogenomics studies.

Chapters 4 to 7 are based on the bioinformatics tools and toxicogenomics studies that were published in collaboration with other researchers in the MARCAR consortium. Chapter 4 describes the ZBIT Bioinformatics Toolbox, a web-based collection of bioinformatics tools for systems biology, expression data analysis, and transcription factor analysis. The chapter provides an overview of the included tools and describes the user interface, the technical setup, and the architecture of the web platform, along with use cases for each major category of the ZBIT Bioinformatics Toolbox. Chapter 5 presents ToxDBScan, a web tool for hepatocarcinogenicity assessment, and provides a detailed mathematical background for the proposed similarity scoring index. In addition, Chapter 5 includes the evaluation of ToxDBScan with data from the MARCAR project, examples of the MOA detection with ToxDBScan, and a proof of concept for a possible extension of the tool with predictive analyses. Chapter 6 introduces two integrative feature types for multiplatform *omics* data, reports a predictive toxicogenomics study which compares single-*omics*, multi-*omics*, and integrated features for hepatocarcinogenicity classification, and discusses the observed results and their implications for the development of alternative testing strategies. Chapter 7 presents MARCARviz, a web platform for interactive visualization of the transcriptomics data that the MARCAR project has generated. The chapter provides a detailed description of the platform architecture and the user interface, as well as a use case for the application of MARCARviz for mechanistic analyses.

To conclude, Chapter 8 discusses the presented results and tools in the context of the MARCAR project and summarizes the conducted research and its potential contributions to preclinical drug and compound development.

Chapter 2

Gene regulation and nongenotoxic carcinogenesis

The regulation of gene expression is a complex, multistage process, which begins with the structural and chemical properties of the DNA molecule and ends with biomolecules that influence and control the life and activity of biological cells. Adverse perturbations of the gene regulation in a single cell can have consequences for the whole organism. The principal example for these errors in gene regulation are tumors, abnormal tissue growths that originate from cells with a defective regulation of the cell's life cycle. Malignant tumors, also called cancer, can spread across the whole organism and are a leading cause of death worldwide. Anand *et al.* (2008) estimate that environmental factors, such as smoking, diet, sun exposure, and environmental pollutants, cause 90-95% of all cancer cases. The formation of tumors can be attributed to three major external factors: physical (e.g., ultraviolet and ionizing radiation), chemical (e.g., tobacco smoke), and biological (e.g., virus infections) carcinogens (IARC, 2014)).

Chemical-induced carcinogenesis is often driven by DNA damaging effects, which cause mutations in genes that regulate the cell life cycle. Genotoxic carcinogens (GCs) are known to induce DNA damage , which initiates tumor formation through DNA mutations. In contrast, nongenotoxic carcinogens (NGCs) promote tumor formation but exhibit no genotoxic effects. While the mechanism of GCs is defined by their DNA damaging properties, NGCs act through a wide range of mechanisms which disturb the regulation of genes that control the proliferation, growth, and death of cells. The characterization and identification of these nongenotoxic effects are the major goals of the research presented in this thesis.

The first section of this chapter illustrates the molecular mechanisms of gene regulation with regard to mRNAs, miRNAs, and proteins. The second section of this chapter describes the *omics* technologies that are used to assess gene regulation in cells. The last section of this chapter illustrates the differences between genotoxic and nongenotoxic carcinogenesis and provides an overview of the mechanisms of NGCs.

2.1 Molecular regulation of gene expression

Gene expression is the process of translating the genetic information that is stored within the chemical structure of the DNA into biologically active proteins. This process is divided into two major steps: (i) the transcription of the genetic code from the stable, two-stranded DNA into a single-stranded mRNA molecule and (ii) the translation of the mRNA molecule into a functioning protein (Twyman, 2001). These two steps can be further dissembled into smaller steps, e.g., the uncoiling of the DNA for transcription, RNA transport, or post-transcriptional modifications of mRNAs. From gene transcription initiation to protein degradation each step is regulated by molecular and biochemical mechanisms. The following sections describe the steps that are relevant for this thesis: the transcription of DNA to mRNA, the post-transcriptional silencing of mRNA by miRNAs, the translation of mRNAs into proteins, and the post-translational modification of proteins. Figure 2.1 shows a overview of these steps of gene regulation. The following sections are based on literature and the book Developmental Biology by Twyman (2001).

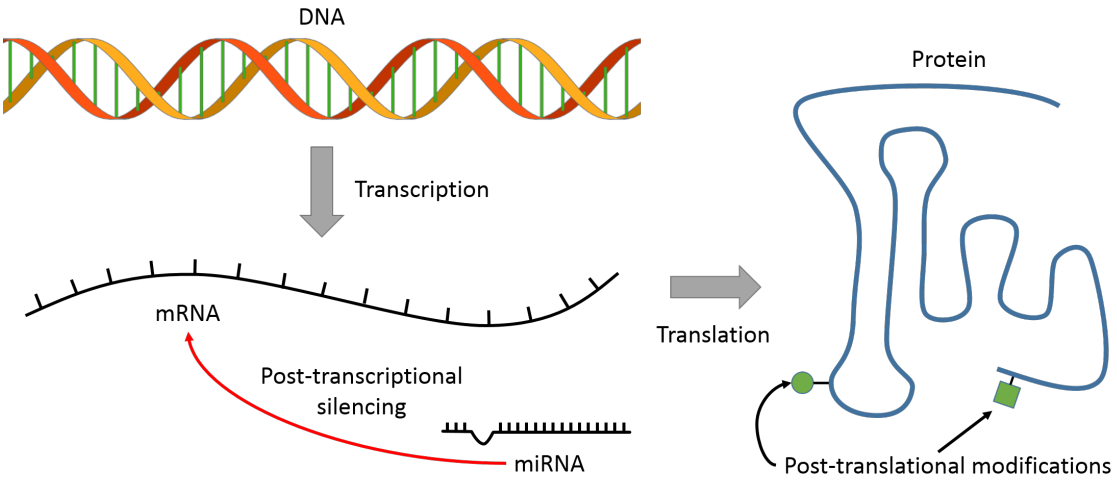


Figure 2.1: **Molecular regulation of gene expression.** The genetic code of the DNA is transcribed into mRNAs, which are the templates for the protein amino acid sequence. The DNA also encodes non-coding RNAs that regulate DNA transcription and mRNA translation. Non-coding miRNAs inhibit mRNA translation by binding to complimentary regions on mRNAs and forming RNA-induced silencing complexes. Protein-coding mRNAs are translated into peptide chains that are cleaved and folded to form the final protein product. Post-translation modifications such as phosphorylation and methylation regulate the activity of proteins.

2.1.1 Transcription of DNA to mRNA

Transcription is the synthesis of RNA molecules based on the DNA template. Transcription is the first and most crucial step of gene expression, as all subsequent steps depend on the transcribed RNA molecules. In eukaryotes, the DNA is transcribed by three RNA polymerases, but only RNA polymerase II produces mRNA for protein synthesis. RNA polymerases I and III transcribe ribosomal and transfer RNAs, which are required to build the ribosome and synthesize proteins. The RNA polymerase II binds to an upstream region of the gene, the promotor. The DNA sequence of the promotor determines the binding affinity of the RNA polymerase II and can vary greatly between genes, which leads to different expression profiles.

In prokaryotes, the ground state of gene expression, i.e., the expression without any transcription altering molecules, is nonrestrictive. In consequence, gene regulation is mainly a question of gene repression. In contrast, the ground state of gene expression in eukaryotes is restrictive. This is the result of the tightly packaged chromatin structure, which prevents the binding of RNA poymerase II to the promotor in the absence of transcription factors that catalyze the binding. The presence or absence of transcription factors is the major driver of differentiation and cell- or tissue-specific gene expression. Transcription factors bind to the promotor to initiate RNA polymerase II activity or interact with enhancer sites on the DNA to increase the speed and stability of the RNA transcription complex. Many transcription factors also influence the chromatin structure to facilitate or impede the access to gene promotors.

Promotor binding and transcription initiation are the dominant regulatory factors of transcription. However, after the transcription has started, the RNA synthesis can be accelerated, slowed, or terminated. For example, RNA elongation (the addition of RNA bases to the already synthesized RNA strand) is regulated by proteins and transcription factors that bind to the RNA transcription complex.

2.1.2 Silencing of mRNA by miRNA

After transcription, the mRNA is subject to post-transcriptional processing. This process of mRNA modification involves capping to improve stability, transport to specific translation locations, and alternative splicing of introns. In addition to these mRNA modifications, mRNAs are also silenced and degraded to maintain steady levels of mRNA for protein synthesis. The abundance of mRNAs in the cell and the translation of mRNAs into proteins is regulated by many factors, such as degradation by non-coding RNAs and enzymes and silencing by miRNAs. These miRNAs form a family of small RNAs of 21-25 nucleotides (nt), which bind to specific target mRNAs and inhibit their translation into proteins (He and Hannon, 2004). The transcriptional regulation of miRNAs is still subject of research (Lee, 2002). Some miRNAs have been found in the introns of both protein-coding and non-coding host genes, which suggests that they are co-regulated with other RNAs (Lagos-Quintana *et al.*, 2003).

By binding at 7 or 8 nt complementary base pair matches in the 3' untranslated region (3'-UTR) of the target mRNA, miRNAs form RNA-induced silencing complexes, which inhibit translation during protein elongation or the release of the protein product (Olsen and Ambros, 1999; He and Hannon, 2004). The post-transcriptional silencing of mRNAs by miRNAs is also shown in Fig. 2.2. The discovery of the imperfect complementarity at the mRNA-miRNA binding sites is the main access point for miRNA target prediction, for which a variety of tools exist (Bartel, 2009). According to Bartel (2009), a simple miRNA target recognition can be performed by a three-step procedure: first, identify the miRNAs seed, which are the nucleotides 2-8 of the 5' region of the miRNA, second, use genome-wide alignments to compile orthologous 3'-UTRs, and third, search these orthologous UTRs for conserved occurrences of the miRNA seed (Bartel, 2009).

The prediction of miRNA targets has revealed that miRNAs regulate a large number of functionally diverse genes (Lim, 2003). In addition, miRNA expression has been found to be stage-specific in development and tissue-specific in different organs (Bartel, 2004). Bartel and Chen (2004) propose that miRNAs act as micromanagement modulator of gene expression by (i) suppressing genes that should not be expressed in a specific cell type and by (ii) adjusting the expression of genes to allow for a tissue-specific gene expression pattern. Lu and Clark (2012) report that genes which are regulated by miRNAs

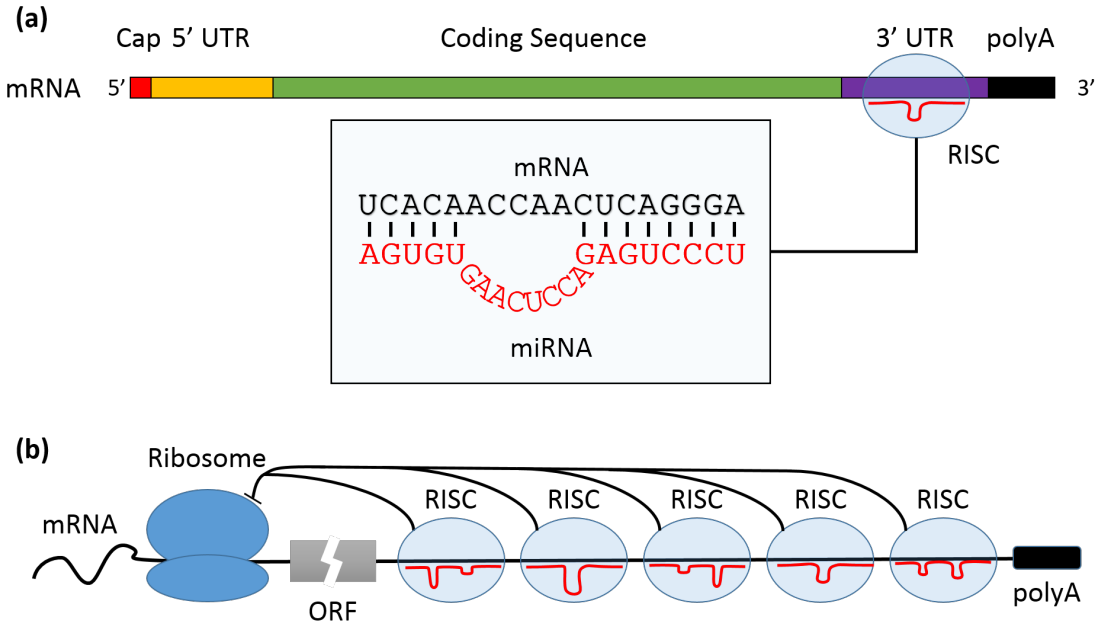


Figure 2.2: **Post-transcriptional silencing of mRNAs by miRNAs.** (a) The miRNA (red) binds to the 3' untranslated region (3'-UTR) of the target mRNA at 7 or 8 nt complementary base pair matches. (b) RNA-induced silencing complexes (RISC) inhibit the elongation at the ribosome and the release of the translated protein.

show a greater variation in expression at the population level compared to genes that are not regulated by miRNAs. Lu and Clark (2012) conclude that post-transcriptional silencing has two major roles in gene regulation: (i) fine-tuning or canalizing the expression levels of a subset of target genes and (ii) promote expression variation of other target genes.

The dysregulation of miRNAs can lead to serious malfunctions of the gene expression regulation. For example, the loss of tumor-suppressive miRNAs leads to an increased expression of cancer-related genes (called oncogenes), whereas the increased expression of cancer-related miRNAs (called oncomirs) can repress tumor repressor genes (Kasinski and Slack, 2011). Targeting these miRNAs has been proposed as a potential cancer therapy and a synthetic mimic of a tumor suppressor miRNA is currently in clinical phase I trials for the treatment of liver cancers (Hayes *et al.*, 2014). The deregulation of miRNAs is also associated with other diseases and the Human microRNA Disease Database lists 572 miRNAs that are involved in 378 diseases (Li *et al.*, 2014).

2.1.3 Translation of mRNA to proteins

Translation is the synthesis of proteins based on the mRNA template and is the second major step in gene expression. The synthesis of proteins is performed by ribosomes, molecular complexes that are formed by ribosomal RNA (rRNA) and proteins. Ribosomes are assembled at the 5' cap of an mRNA and start the protein synthesis when encountering an AUG codon. The protein is elongated with amino acids carried by transport RNAs (tRNAs), which match the codon at the current position of the ribosome. When a stop codon is detected, release factors cause the ribosome to disassemble and release the synthesized protein.

Like all stages of gene expression, translation is regulated by multiple molecular and structural mechanisms. In eukaryotes, translation initiation is controlled by enzymatic factors that catalyze ribosome binding (Kozak, 1999). In addition, mRNA binding factors affect the translational regulation, e.g., miRNAs, which repress protein elongation and release. The number of ribosomes is also a limiting factor for the amount of proteins produced by a cell.

The translation is one of the main targets of antibiotics due to the evolutionary differences between prokaryotic and eukaryotic ribosomes. Most antibiotics that are used clinically inhibit either the binding of tRNAs to the ribosome or the translocation of the ribosome along the mRNA (Wilson, 2013). The disruption of translational control has also been linked to cancer formation, e.g., by selective translation of oncogenes, and has been found to be specific for various types of cancer or disease stages (Silvera *et al.*, 2010).

2.1.4 Post-translational modification of proteins

After the release from the ribosomal complex, most proteins undergo post-translational modifications (PTMs) that alter the protein structure, the amino acid composition, or properties of the amino acid side chains. The most relevant PTMs for this thesis are chemical modifications of the amino acid residues, such as phosphorylation, acetylation, and methylation.

The reversible phosphorylation of serine, tyrosine, and threonine residues in proteins is a major regulatory mechanism of protein activity (Johnson and Lewis, 2001). In humans, approximately 86.4% of serine, 11.8% of threonine, and 1.8% of tyrosine residues are phosphorylated (Olsen *et al.*, 2006). Although tyrosine phosphorylation is less common, it is critical for many cell signaling processes. Several oncoproteins, growth factors, and hormones show tyrosine phosphorylating activity (Johnson, 2009b). Phosphorylation can change the function of proteins through multiple mechanisms. The negative charge of the phosphoryl group can lead to the formation of bonds with arginine residues, which results in conformational changes of the protein (Mandell *et al.*, 2007). In other cases, enzyme activity is inhibited without conformational changes, e.g., by the phosphate group acting as a steric blocking agent (Russo *et al.*, 1996) or impeding substrate recognition (Hurley *et al.*, 1990). The transfer of phosphate groups to amino acid residues is catalyzed by proteins that are called kinases. The identification of the human kinome, i.e., the set of all kinases, led to the discovery of seven major groups of kinases (Manning *et al.*, 2002). Due to their prominent role in the regulation of cellular signaling, protein kinases have been studied extensively and were linked to many types of cancers and diseases. In 2008, ten kinase inhibitors had been approved as drugs for the treatment of several cancer types, and several more are currently in clinical trials (Johnson, 2009a).

The acetylation of proteins usually occurs at lysine residues and neutralizes the positive charge of the lysine, which changes the protein function (Choudhary *et al.*, 2009). Histone acetylation plays an important role in the regulation of gene expression and has long been known to increase transcription (Allfrey *et al.*, 1964). More recently, lysine acetylation has been shown to influence *p53* function and interactions (Yang and Seto, 2008), which suggests a broader role of protein acetylation in gene regulation and protein function. Acetylation and deacetylation are catalyzed by acetyltransferases and deacetylases, both for histones and lysine residues in general (Shahbazian and Grunstein, 2007). Acetyltransferases and deacetylases have been identified as potential targets of drugs for cancer or neurodegenerative diseases and are currently tested in clinical trials (Choudhary *et al.*, 2009). However, despite the recent increase in interest, the knowledge of *in vivo* acetylation sites is still sparse.

Protein methylation has long been associated mainly with histones and chromatin modification, where they are a major part of the "histone code", which regulates gene transcription (Biggar and Li, 2014). Methylation occurs predominantly at lysine and arginine residues, which can be methylated up to two (arginine) or three (lysine) times. Recently, the methylation of non-histone proteins has been found to have a regulatory

role in many cellular processes (Paik *et al.*, 2007). Similar to the role of kinases for phosphorylation, methyltransferases that are specific for arginine or lysine act as an activator by transferring methyl groups to the amino acid residues, which leads to the binding of methyl-binding proteins or alters protein structure and activity. After the activating stimulus is withdrawn, demethylases erase the methyl group and restore the previous protein state. Methylation has long been considered to be a stable modification, but the discovery of non-histone protein methylation has also shown that methylation can be as dynamic as phosphorylation events (Dhami *et al.*, 2013) and is linked to several crucial cell signaling pathways, e.g., the mitogen-activated protein kinase (MAPK) signaling cascade (Mazur *et al.*, 2014). The SMYD3 gene in the MAPK signaling cascade also links lysine methylation to cancer (Mazur *et al.*, 2014), which indicates that methylation also regulates oncogene expression.

2.2 Measuring gene regulation with *omics* technologies

Gene expression profiling is of great use for toxicological studies because changes in the mRNA abundance are among the first detectable responses to compound administration. The earliest methods for mRNA transcript profiling such as the Northern blot by Alwine *et al.* (1977) performed one assay for each transcript. Meanwhile, these traditional methods have been replaced by techniques that can quantify the expression of thousands of genes simultaneously. Among the most used technology in larger projects are mRNA microarrays, which are commercially available since the 2000's (Bumgarner, 2013). Despite the recent advent of next-generation sequencing for transcript profiling (RNA-seq), microarrays remain an important technology for gene expression profiling because they are cheaper, require less data storage, and provide well-established analysis pipelines (Zhao *et al.*, 2014). Besides mRNA arrays, manufacturers provide microarrays for a wide range of other biomolecules such as non-coding RNAs, DNA methylation, and proteins. This section describes the three types of array-based technologies that we used to generate the data for the research presented in this thesis: Affymetrix mRNA microarrays, Agilent miRNA microarrays, and reverse phase protein arrays (RPPAs).

2.2.1 Gene expression microarrays

Gene expression microarrays, also called DNA or mRNA microarrays, have been applied in many fields of biology, such as pharmacogenomics (Debouck and Goodfellow, 1999), cancer research (Khan *et al.*, 1999), and pathology (Becich, 2000). The application of microarrays involves multiple steps: array fabrication and probe spotting, hybridization of probes and sample material, hybridization signal detection, and data analysis. Microarrays are fabricated on glass, silicon, or plastic surfaces and spotted with DNA probes, i.e., synthetic oligonucleotides, by in situ synthesis. The extracted DNA material is labeled with reporters (fluorescent dyes) and a whole array scanner measures the signal

of the bound reporter probes. Thus, an image of the microarray contains the complete hybridization pattern and can be used for the hybridization analysis. The high-density arrays manufactured by Affymetrix utilize a massively parallel approach for DNA hybridization, with hundreds of thousands of test sites (probes) in less than 2 cm^2 (Lockhart *et al.*, 1996).

To profile mRNA abundance in samples, the RNA is extracted, isolated, and purified with RNA preparation kits and RNA quality is assessed. The extracted RNA is labeled with a fluorescent dye that can be detected by the scanner and hybridized to the microarray probes. Affymetrix provides reagent kits to perform the necessary hybridization procedure, which includes hybridization, washing and staining. After hybridization, the scanner captures a fluorescence image and calculates a signal intensity for each probe. The probe signal intensity is a function of the amount of sample RNA that has hybridized to a probe. These intensities are the raw data input for the bioinformatics analysis, which is described in Chapter 3.

2.2.2 MiRNA microarrays

Perturbed miRNA expression can lead to dysfunctional cell regulation and is specific for different tissues, diseases, and cancers. For this reason, miRNA profiling is of great interest and microarray manufacturers have adapted the mRNA microarray technology to miRNAs during the last decade. The main challenge in miRNA profiling is the small size of miRNAs ($\sim$ 22 nt) and the high degree of homology between different miRNAs (D'Andrade and Fulmer-Smentek, 2012). Commercially available mRNA arrays typically use probes of 25 nt (Affymetrix) or 60 nt (Agilent), which necessitates a different probe design for miRNA microarrays. The presence of unspliced precursor miRNAs must also be considered in the probe design (D'Andrade and Fulmer-Smentek, 2012). The miRNA data in this thesis was generated with Agilent miRNA microarrays, which use a labeling method and probe design that is optimized for miRNAs (Wang *et al.*, 2006). Agilent also provides labeling and hybridization kits for the miRNA microarrays to standardize sample processing. Subsequently, the miRNA microarrays are measured with commercial scanners and images and signal intensities are analyzed with bioinformatics methods.

2.2.3 Reverse phase protein arrays

Proteins are biologically active molecules that perform most cellular functions such as catalyzing chemical reactions, transporting other molecules, or replicating the DNA. Measuring protein expression and activation states is the most direct assessment of the biological state of a cell. Traditional methods (e.g., two-dimensional gel electrophoresis in combination with mass spectrometry) identify mostly high abundance proteins (Gygi *et al.*, 2000). However, up to 90% of the proteins in a cell are present in low copy numbers (Miklos and Maleszka, 2001) and there is no protein amplification method available.

In addition, protein immobilization and binding is difficult due to the large differences in function and structure. Proteins show high diversity, including anti-bodies, water-soluble enzymes, and water-insoluble membrane proteins.

RPPAs utilize microarray technology to profile protein expression with specific antibodies, which function as protein-binding probes. In contrast to most microarray technologies, in RPPAs the samples are immobilized on the chip instead of the probes. The chip is spotted with beads of the sample solution and nonspecific binding sites on the array are blocked. Then, the chip is stained with a protein-specific antibody that is labeled with a fluorescence marker, incubated, and washed to remove unbound antibodies. A fluorescence detection system takes an image of the chip and measures the fluorescence signal intensity for each spotted sample. Depending on the availability of specific antibodies, RPPAs can also detected protein modifications, e.g., phosphorylation at functionally relevant sites, which allows the profiling of the activity state of proteins.

RPPAs have been applied in a wide range of applications, e.g., to identify risk factors for liver diseases (Morales-Ibanez *et al.*, 2016). We have also combined RPPAs with RNA microarrays for the integrated assessment of the effects of the NGC phenobarbital in the mouse liver (Braeuning *et al.*, 2016). To generate the data used in this thesis, we used the ZeptoMARK system, which uses a specialized fluorescence reader called ZeptoREADER (Pawlak *et al.*, 2002).

2.3 Mechanisms of chemical-induced carcinogenesis

Carcinogenesis is the formation of cancer, i.e., malignant tumors, which are characterized by uncontrolled cell proliferation and the ability to spread throughout the organism. DNA damage, either naturally occurring or through endogenous and exogenous factors, is considered to be the leading cause of cancer formation (Bernstein *et al.*, 2013). The causes of DNA damage can be roughly categorized to be either physical, such as ultraviolet and ionizing radiation, biological, e.g., infections with viruses, or chemical (IARC, 2014). However, long-term rodent carcinogenicity studies have repeatedly demonstrated that chemicals that show no genotoxic activity can also induce carcinogenesis. A study of approved pharmaceuticals found that 50% of the pharmaceuticals were rodent carcinogens but only six were found to be genotoxic (Van Oosterhout *et al.*, 1997). This demonstrates that many approved pharmaceuticals are NGCs, which are carcinogenic in rodents but show no DNA damaging potential. NGCs act through a range of different mechanisms, in contrast to GCs, for which the carcinogenic profile is determined by direct DNA damage (Fig. 2.3). The following section describes genotoxic carcinogenesis and illustrates proposed mechanisms of NGCs.

2.3.1 Genotoxic carcinogenesis

Genotoxic carcinogens induce cancer formation by causing DNA damage through direct interactions with the DNA. For this reason, they are also known as DNA-reactive carcinogens. The DNA reactivity can occur either by the carcinogen itself or by DNA-reactive metabolites of the carcinogen. In contrast, NGCs act through secondary mechanisms without directly damaging the DNA. However, the distinction between GCs and NGCs is not absolute, as many GCs can also act through mechanisms other than DNA damage and NGCs may indirectly lead to DNA damage (Benigni *et al.*, 2013). Substances that change the DNA sequence are only a subset of genotoxins, as genotoxins can also cause DNA lesions, i.e., sites where the structure or the base-pairing of the DNA are disrupted. Types of DNA lesions include sites where one base is missing (abasic sites), single- and double-strand breaks, and covalent links between the two DNA strands (interstrand crosslinks). DNA lesions can have strong effects on cells by inhibiting DNA replication and transcription and require DNA repair, which is error-prone and can lead to mutations.

Most GCs or their metabolites are strong electrophilic reactants, which bind to nucleophilic sites in the DNA, RNAs, and proteins. The formation of carcinogen-DNA adducts is linked to the occurrence of mutations. Altered bases are fixated by erroneous

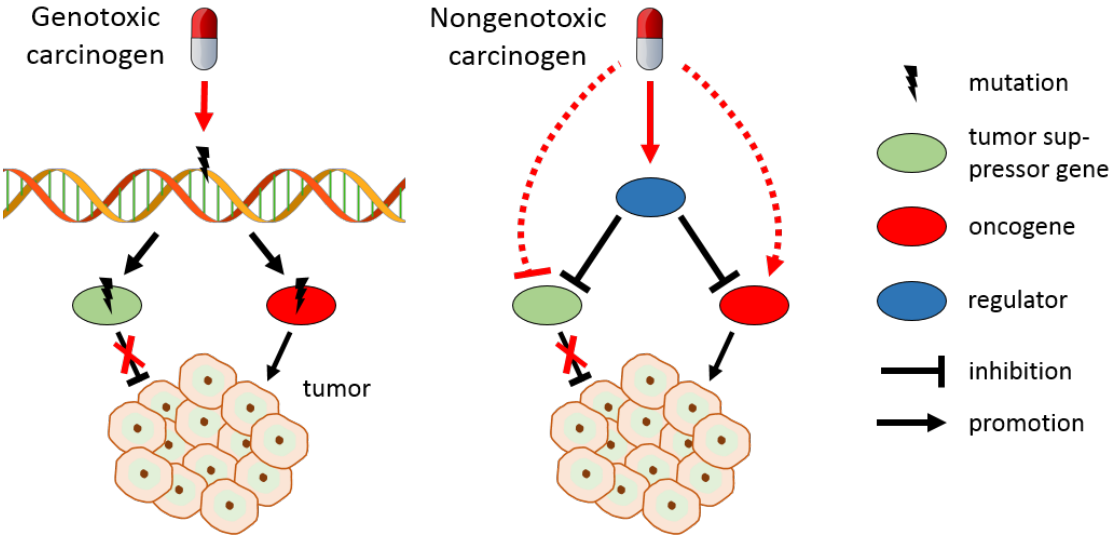


Figure 2.3: **Genotoxic and nongenotoxic tumorigenesis.** Genotoxic carcinogens can bind to the DNA to form DNA adducts and cause DNA mutations. Mutated tumor suppressor genes or oncogenes can lead to tumor formation. In contrast, nongenotoxic carcinogens perturb the regulation of oncogenes and tumor suppressor genes by altering the function of regulatory genes, which promotes tumor growth. Red arrows indicate direct effects of the carcinogen, black arrows indicate indirect effects.

base-pairing during DNA replication to form potentially initiated cells (Miller and Miller, 1981). Some GCs require metabolic activation to form electrophilic metabolites that can bind to the DNA and form DNA adducts. For example, most N-nitroso compounds, such as N-nitrosomethylurea or dimethylnitrosamine, require metabolic activation (Magee, 1969). The oxidative removal of one N-alkyl group produces a monoalkylnitrosamine, which is unstable and rearranges to form electrophilic ions. Well-known GCs include aflatoxin B1, a food contaminant that is a major risk factor for liver carcinomas (Kew, 2003), components of tobacco smoke (e.g., nitrosamines and polycyclic aromatic hydrocarbons, Hecht (1999)), and ethanol, which is metabolized to several DNA damaging metabolites, e.g., acetaldehyde and reactive oxygen species (Boffetta and Hashibe, 2006).

Most GCs are not species-specific because DNA mutations are a general initiating factor in carcinogenesis. For this reason, a battery of bacterial and *in vitro* assays has been developed for the identification of substances with genotoxic effects. One regularly used genotoxicity assay is the Ames test, a bacterial reverse mutation assay (Ames *et al.*, 1975). The Ames test detects genotoxins which cause point mutations and frameshifts. As the Ames test requires only a few days, it is a standard assay for the risk assessment of new chemicals. The Ames test has also been shown to identify a high percentage of known carcinogens (McCann *et al.*, 1975).

The majority of known chemical carcinogens is genotoxic, and the two largest risk factors for lifestyle associated cancer are tobacco smoking and alcohol, which both act as GCs (Anand *et al.*, 2008; Hernández *et al.*, 2009). The accumulation of activating mutations in tumor-promoting oncogenes and silencing mutations in tumor suppressor genes is a critical feature of carcinogenesis. However, Hernández *et al.* (2009) report that 45 of the 371 chemicals (12%) considered to be carcinogenic to humans (IARC Group 1) or likely to be carcinogenic to humans (IARC Groups 2A and 2B) are not genotoxic and thus qualify as human NGCs. In consequence, a risk assessment driven purely by genotoxicity findings is not sufficient, and nongenotoxic carcinogenesis has to be considered in the development of new chemicals and drugs

2.3.2 Nongenotoxic carcinogenesis

NGCs are capable of promoting cancer formation without direct interaction with the DNA. In contrast to the MOA of GCs, where the induced DNA damage leads to mutations in tumor suppressor or oncogenes, the knowledge of the mechanisms of tumor induction and promotion of NGCs is sparse. NGCs show a wide range of tumor-inducing and tumor-promoting modes of action, such as peroxisome proliferation, immunosuppression, endocrine modification, or cytotoxicity (Williams, 2001).

Melnick *et al.* (1996) argued that limiting the evaluation of NGCs to their nongenotoxic effects is not sufficient, as some NGCs have been found to form DNA adducts and induce DNA mutations although they are Ames-negative. For example, tamoxifen, a drug used in the treatment of breast cancer, was long thought to be a rat-specific, nongeno-

toxic liver carcinogen due to negative *in vitro* tests for mutagenicity and an assumed hormonal perturbation mechanism (Tucker *et al.*, 1984). However, it was later shown that tamoxifen is activated by rat and human liver microsomes to form DNA adducts (Pathak and Bodell, 1994), which provided evidence for a human relevance of tamoxifen induced liver cancers. In 2012, tamoxifen has been reevaluated by the IARC and was classified as a Group 1 carcinogen with sufficient evidence that tamoxifen is a carcinogen in humans (IARC, 2012). The argument by Melnick *et al.* (1996) demonstrates that the classification into GCs and NGCs is not a clear distinction, as NGCs may initiate tumors through weak genotoxicity and GCs may promote initiated cells to form tumors by nongenotoxic mechanisms (Benigni *et al.*, 2013). Mostly, the term nongenotoxic is used to mean Ames negative because the Ames test is the most common assay to determine the genotoxic potential in chemical development. Accordingly, the term nongenotoxic is used to mean Ames negative in this thesis.

Genotoxic substances can be identified by a range of short-term bacterial and *in vitro* tests and are excluded from further development. In contrast, NGCs are negative in these genotoxicity assays and are only detected in the lifetime rodent bioassays (LRB, Hernández *et al.* (2009)). However, the LRB is time-consuming, expensive, and no longer in line with new policies for the reduction of animal testing. For this reason, several projects and consortia have investigated NGCs in the last decade, among them the MARCAR project. These projects try to expand the knowledge of the mechanisms of nongenotoxic carcinogenesis to provide cancer assays that require fewer animals and less time than the LRB. For human cancer risk assessment, the LRB is complemented with data from 90-day toxicity studies and toxicological studies of chemical kinetics and disposition in rodents (Hernández *et al.*, 2009). Human epidemiological data is also integrated into risk assessment when available. As NGCs act through a variety of mechanisms, the following sections provide an overview of some of the most prevalent mechanisms of human nongenotoxic carcinogenesis based on the reviews by Hernández *et al.* (2009) and Silva Lima and Van der Laan (2000).

Endocrine modifiers

The glands that secrete hormones into the blood constitute the endocrine system. Endocrine modifiers induce or promote cancer mainly by modifying the levels of hormones or binding to hormone receptors. For example, 17β-estradiol and its metabolites can bind to the estrogen receptor, which induces strong mitogenic effects by altering pathways mediated by the estrogen receptor. Estrogen is assumed to be a major factor in cancer induction because it modulates transcription factors that regulate proliferation, differentiation, and apoptosis (Chen *et al.*, 2008). Hormones that bind to the estrogen receptor are considered human NGCs (IARC, 1987). Other hormone receptors that are targets for human NGCs are the progesterone receptor, the aryl hydrocarbon receptor, and the thyroid hormone receptor. Some nongenotoxic substances can bind to multiple receptors, for example, the pesticide dichlorodiphenyltrichloroethane (DDT). DDT

promotes hormone-dependent pathology, induces several cytochrome P450 genes, and is associated with increased incidences of lymphoma and lung cancers in workers exposed to the pesticide (Sierra-Santoyo *et al.*, 2000). These multi-receptor binding substances can further complicate the human risk assessment of compounds. Ultimately, the imbalance of the hormonal control by estrogen, progesterone, or other hormones, coupled with cytotoxicity and other effects of NGCs is thought to be a major driver of tumor promotion.

Goitrogens, i.e., substances that interfere with iodine uptake and thus disturb the production of thyroid hormones in the thyroid gland, are the largest group of non-receptor mediated endocrine modifiers. This class includes anti-thyroid drugs as well as central nervous system-acting drugs (e.g., phenobarbital), chlorinated hydrocarbons (e.g., chlordane), and others (Hernández *et al.*, 2009). These substances increase the peripheral production of thyroid hormones and the secretion of thyroid-stimulating hormone, which have been linked to the induction of cell hypertrophy and hyperplasia and thyroid tumors in the LRB.

Cytotoxicity and chronic cell injury

In tissues where tumor incidences are found in LRBs, cytotoxicity and regenerative hyperplasia are often observed after the administration of NGCs (Dietrich and Swenberg, 1991). The regenerative proliferation response to chronic cell injury leads to enhanced cell replication, which is associated with cancer initiation and promotion. Melnick *et al.* (1996) suggested that this is due to the reduced time available for DNA repair mechanisms during cell division, which increases the probability of the fixation of endogenous or exogenous DNA damage. Examples for NGCs with a proposed cytotoxic MOA are the arsenic compounds dimethylarsinic acid and monosodium methane arsenate. These two also inhibit DNA repair and suppress *p53*, which supports the hypothesis that premature replication fixates DNA damage in exposed cells (Cohen *et al.*, 2006; Salnikow and Zhitkovich, 2008). In general, tumor promotion by inhibiting DNA repair and inducing cell cycle progression is often observed for cytotoxic NGCs (Safe, 1989; IARC, 1999; Stickney *et al.*, 2003).

Inhibition of gap junction intercellular communications

The maintenance of homeostasis, i.e., a stable state of the internal cell conditions, requires intercellular communication. Gap junctions, which are plasma membrane channels that are formed by proteins, link adjacent cells and enable the exchange of small molecules (Kumar and Gilula, 1996). This communication between cells is important for the control of cell differentiation and proliferation. The disturbance of intercellular communication by the inhibition of gap junctions has been observed for many NGCs such as chlordane, DDT, phenobarbital, arsenic compounds, and peroxisome proliferators (Hernández *et al.*, 2009; Rivedal and Witz, 2005; Budunova and Williams, 1994;

Cowles *et al.*, 2007). The exact mechanisms of the inhibition of gap junction communications are not clearly established for most NGCs. Cowles *et al.* (2007) found that gap junctions can be affected by both direct (by DDT or Wy-14,643) and indirect (by tetrachloride) connexin inhibition, particularly connexin 32 in the liver of rats. This also demonstrates that NGCs often act through multiple of the presented MOAss.

Other mechanisms

In addition to the presented mechanisms, there are several others that have been proposed to contribute to carcinogenesis, e.g., peroxisome proliferation, immunosuppression, or oxidative stress (Hernández *et al.*, 2009). However, several nongenotoxic substances act through unique and not fully understood mechanisms. For example, carbon tetrachloride is activated by CYP genes to form trichloromethyl radicals, which form DNA adducts and disrupt cellular processes by inducing tumor necrosis and promoting growth factors (Weber *et al.*, 2003). Perchloroethylene (PCE) is a solvent that is used by the dry cleaning industry and is a nongenotoxic rodent carcinogen which is suspected to be less relevant to humans. PCE is metabolized to trichloroacetic acid, which induces hepatocellular peroxisomes in rodent livers (Wernke and Schell, 2004). In general, peroxisome proliferators, such as Wy-14,643, are suspected to be rodent-specific, although the exact mechanism of carcinogenesis is not entirely understood.

Overall, NGCs are known to act through a wide range of mechanisms and often multiple mechanisms are involved. As NGCs cannot be detected by the Ames test and require extensive mechanistic and toxicological investigation, they remain an important factor in human risk assessment during the development of new chemicals and drugs. For this reason, toxicogenomics has gained much momentum during the last decade, and several large projects and consortia have started to investigate new methods for the early identification of NGCs. The next chapter provides an introduction to toxicogenomics and the statistical and machine learning techniques that have been proposed as new short-term alternatives to the LRB.

Chapter 3

Introduction to toxicogenomics and machine learning

Toxicogenomics uses automated high-throughput techniques, statistical and machine learning methods, and computational models to address toxicological research questions. The maturation of mRNA microarray technology provided toxicologists with powerful new instruments for measuring the effects of chemicals on living organisms. More recently, the advent of next-generation sequencing such as RNA-seq further enhanced the repertoire of tools for assessing adverse drug effects. These new technological opportunities for toxicologists and life science researchers initiated the development of new bioinformatics methods that were needed to analyze the increasing amounts of data that current experiments produce. Databases of microarray data are growing fast, but batch effects due to different protocols in different groups interfere with a simple combination of datasets, which necessitates normalization and methods to account for batch effects. Also, researchers and manufacturers are adding more *omics* layers to the repository of high-throughput technologies. Mass spectrometry and RPPAs measure protein abundance, DNA methylation arrays generate epigenetic maps of DNA modification, and new, global expression microarrays include non-coding RNAs with regulatory functions. To make full use of multilayered *omics* data, researchers need methods that can integrate these different types of molecular data and extract meaningful biological signals.

Whereas mechanistic toxicogenomics uses *omics* data to examine a single substance or a particular metabolic pathway, predictive toxicogenomics attempts to construct computational models and signatures that recognize groups of compounds with similar adverse effects. Pharmaceutical companies and regulatory agencies explore predictive toxicogenomics as an economical alternative to long-term animal tests such as the LRB, which are expensive and time-consuming.

The first section of this chapter describes the bioinformatics methods that we used to process and analyze the *omics* data measured with microarray technologies. The second section provides an overview of the machine learning methods and validation strategies that we used in our predictive toxicogenomics studies. The last section reviews the current state of toxicogenomics and presents several studies that are relevant to this thesis.

3.1 *Omics* data processing and analysis

The readouts of the microarray systems presented in Chapter 2 are fluorescence signal intensities. These raw signal intensities are subject to systematic, instrumental, and random noise, which introduces a bias that needs to be removed before the signal intensities can be used to perform quantitative analyses. Microarray data processing encompasses quality control to detect microarrays that did not function correctly and raw data normalization to eliminate systematic bias and make samples comparable. After the processing, statistical and numerical methods are applied to identify affected biomolecules and determine the effect strength. This section first provides an overview of quality control and normalization methods, followed by an outline of the statistical and numerical methods used for data analysis.

3.1.1 Microarray quality control

Quality control is an essential step for microarray data processing. Whereas systematic and instrumental bias or batch effects can be addressed by experimental design and normalization, random noise in the spotting and hybridization process or experimental errors cannot be removed automatically. For example, spatial bias due to contaminations or sample handling mistakes that affect amplification and hybridization can confound experiments and introduce effects that are not a result of drug administration. Schuchhardt *et al.* (2000) list fluctuations that can affect all steps of the probe, sample or array preparation, hybridization, and signal detection. Examples of fluctuations that need to be detected during quality control are failed PCR amplification, unequally distributed sample solution, hybridization failure due to temperature, time, or buffering conditions, and overshining of signals due to external factors.

Several software packages are available to detect these experimental artifacts and remove affected arrays from further analyses. For the MARCAR data, we used the `arrayQualityMetrics` package by Kauffmann *et al.* (2009), which is generic and can be applied to many types of microarrays independent of manufacturer. Specialized software for Affymetrix (Parman *et al.*, 2016), Illumina (Dunning *et al.*, 2007), or two-color cDNA arrays (Buness *et al.*, 2005) are also available. The `arrayQualityMetrics` package provides several analyses for detecting potential outliers such as density plots, principal component analysis (PCA), pairwise distance matrices, or false color plots. Figure 3.1 shows examples of these common quality control plots that were generated with `arrayQualityMetrics`. We used both visual inspection methods (e.g., false color plots and PCA plots) and statistical cutoffs (e.g., pairwise distances and dynamic range) to identify and remove microarrays of insufficient quality from the analysis.

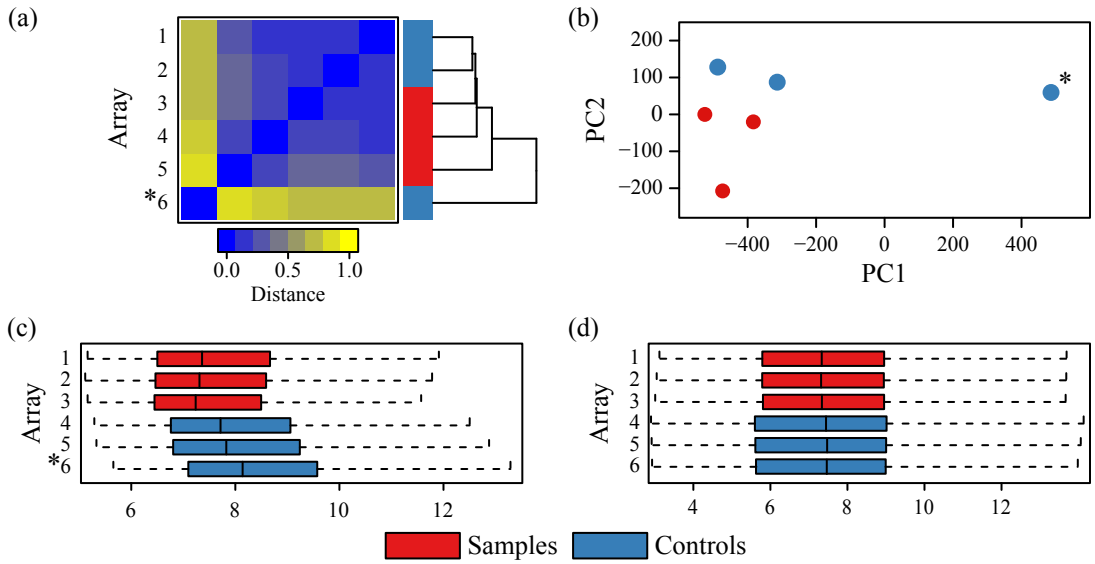


Figure 3.1: **Microarray quality control and normalization.** (a) Pairwise distance heat maps. (b) PCA identifies arrays that show global deviations, possibly due to processing errors. The asterisk marks the outlier. (c) Array density plots identify arrays with perturbed overall intensity distributions. (d) Array density plot after normalization. Intensities are normalized to have similar distributions.

3.1.2 Normalization methods for microarrays

Normalization is the scaling of microarray readouts to make data from multiple arrays comparable (see Fig. 3.1(d), Hill and Whitley (2003)). As stated before, microarray experiments are influenced by systematic and stochastic fluctuations, which can introduce noise into data and confound the analysis (Schuchhardt *et al.*, 2000). Sources of these variations between samples include varying sensitivity to RNA degradation, systematic fluctuation of the labeling, small differences in hybridization parameters (e.g., temperature, time, or sample volume), nonspecific binding, and saturation effects (Schuchhardt *et al.*, 2000). For two-color microarrays, the two conditions that are compared can be considered to be comparable because they were subject to the same preparation, hybridization, and detection process. In contrast, for one-color microarrays the raw signal intensities need to be normalized before they can be compared. To eliminate the noise in the data, systematic fluctuations are measured by control probes or statistically estimated.

Quantile normalization

Amaratunga and Cabrera (2001) introduced quantile normalization for the analysis of viral DNA microarrays, which achieves a robust readout standardization (Bolstad *et al.*, 2003). Quantile normalization standardizes the distribution of probe intensities for all arrays in an experiment.

The procedure for quantile normalization is described by Bolstad *et al.* (2003) as follows: (i) build a matrix X of dimension $p \times n$, given n arrays of with p probes, (ii) sort all n columns of X to get X_{sort}, (iii) assign the mean across rows in X_{sort} to each element in a row to get X'_{sort}, and (iv) rearrange the values in each column to revert X'_{sort} to the original order in X, which gives $X_{\text{normalized}}$. Thus, quantile normalization represents the transformation $\mathbf{x}'_i = F^{-1}(G(\mathbf{x}_i))$, where $\mathbf{x}_i$ is the readout of array i, G is the empirical distribution of each array, and F is the empirical distribution of the averaged sample quantiles (Bolstad *et al.*, 2003). The distribution F can also be replaced by an arbitrary reference distribution, such as the normal distribution. Quantile normalization is a complete data method, i.e., it combines information from all arrays in the normalization transformation.

Quantile normalization is a generic normalization algorithm that can be applied independent of the array manufacturer. We used quantile normalization to standardize RPPA readouts. For commercially distributed microarrays, specialized normalization software and methods are available that account for specific microarray designs.

Robust multi-array average normalization

Affymetrix microarrays include perfect match (PM) and mismatch (MM) probes to estimate noise by nonspecific binding. The normalization provided by the Affymetrix Microarray Suite deducts the MM probe signals from the corresponding PM probe signals to calculate the probe set signal (Hill and Whitley, 2003). However, Irizarry *et al.* (2003) report that MM probes may be detecting both true signal and nonspecific binding. Thus, estimating the true signal as the difference between PM and MM probe signal adds noise without removing bias, whereas estimaing the true signal by dividing the PM signal by the MM signal results in a biased signal. For this reason, Irizarry *et al.* (2003) propose robust multi-array average (RMA) normalization, which discards all MM probe signals and estimates the true signal with a linear additive model that is based on three major observations. First, PM probe signals grow approximately linear to concentration, second, signal variance is roughly constant, and third, probe-specific affinity is approximately additive. Irizarry *et al.* (2003) also adjust the model to account for nonspecific binding and background noise. The log scale measure of expression μ_{in} for array i and probe set n is estimated with a linear additive model:

$$Y_{ijn} = \mu_{in} + \alpha_{jn} + \varepsilon_{ijn}, i \in \{1,\ldots,I\}, j \in \{1,\ldots,J\}, n \in \{1,\ldots,N\}$$

where I, J, and N are the numbers of arrays, probes, and probe sets, Y_{ijn} is the log-transformed, background-corrected, quantile-normalized PM probe signal, α_{jn} is the probe affinity effect, and ε_{ijn} is an independent, identically distributed error term with mean 0 (Irizarry *et al.*, 2003). The background correction uses an additive model, where the observed intensity S is the sum of signal X and background noise Y. Irizarry *et al.* (2003) assume that X follows an exponential distribution, and Y follows a normal

distribution, with $Y \geq 0$, which allows estimating the background-corrected signal by $E(X|S = s)$ (full details are available from Bolstad (2004)).

The RMA algorithm was originally motivated by observations of Affymetrix microarray data but has been modified and applied to microarrays from other manufacturers. For Agilent microarrays, López-Romero *et al.* (2010) show that RMA normalization without background correction is equally precise as other methods that are designed specifically for Agilent arrays. We used the RMA implementation in the `affy` package (Gautier *et al.*, 2004) to normalize Affymetrix microarray data and the modified RMA algorithm in the `AgiMicroRNA` package (López-Romero, 2011) to normalize Agilent microarray data.

3.1.3 Statistical analysis of microarray data

The analysis of data from high-throughput microarray experiments requires efficient and robust statistical methods. This section focuses on experimental designs that include a reference condition (called control) for each sample. The sample-control design is the most common experimental design because it addresses some common issues of microarray data analysis. First, expression values are semi-quantitative, i.e., they represent not the absolute transcript abundance in a sample but an experiment-specific relative transcript abundance. Second, laboratory-, time-, or technician-specific batch effects may confound the biological effects even after normalization (Leek *et al.*, 2010). The inclusion of controls allows the estimation of the effects of a specific treatment or condition compared to a reference condition and also the assessment of batch effects. For example, in the MARCAR toxicogenomics studies, the reference condition was treated with the compound vehicle (i.e., the solvent used to dilute the compound) that was also used for the compound-treated animals to eliminate confounding effects of the vehicle substance. In consequence, several conditions, e.g., animals treated with different carcinogenic compounds, can be compared based on the changes in transcript abundance relative to the respective reference. Batch effects can be removed either by using batch-matched controls that are processed in the same batch (i.e., at the same laboratory, the same time, and by the same person) or by randomly distributing conditions and controls across all batches, which allows a statistical estimation of the batch effect (Chen *et al.*, 2011). All studies that are described in this thesis use batch-matched controls to avoid batch effects and provide reference-based estimates of the condition-specific effects on transcript abundance.

Experimental designs with replicates allow the assessment of both the power of an effect and its significance. Effect power is measured by the fold change and the significance is commonly expressed by p-values. The following paragraphs describe the calculation of fold changes and p-values with the `limma` package for R/Bioconductor Smyth (2005), which we used in this thesis.

Fold change

The fold change is a semi-quantitative measure of transcript abundance and is calculated as the ratio of average normalized signal intensities in samples and controls. The fold change x_g is calculated as

$$x_g = \frac{\sqrt[n]{\prod_{i \in 1,\ldots,n} s_{ig}}}{\sqrt[m]{\prod_{j \in 1,\ldots,m} c_{jg}}}$$

where s_{ig} and c_{jg} are the normalized signal intensity of gene g in sample s_i and control c_j. Because signal intensities are ratios of detected fluorescence, the geometric mean must be used. However, most normalization methods, including the RMA algorithm, use a log-transformation and thus report logarithmized expression values. Then, the fold change is calculated as the difference between the arithmetic means of logarithmized expression values in samples and controls:

$$x_g = \frac{\sum_{i \in 1,\ldots,n} \log(s_{ig})}{n} - \frac{\sum_{j \in 1,\ldots,m} \log(c_{jg})}{m}$$

The MicroArray Quality Control (MAQC) project presents strong evidence that a gene ranking by fold change is reproducible across microarray platforms (Guo *et al.*, 2006; Patterson *et al.*, 2006; Shi *et al.*, 2005). Shi *et al.* (2006) argue that a combination of fold change ranking and non-stringent significance filtering produce more reproducible gene lists than ranking and filtering genes based on t-test statistic alone. Also, Patterson *et al.* (2006) demonstrated that the fold change ranking is less dependent on the used background correction and scaling method.

Nevertheless, when only considering fold changes, weak but significant effects might be overlooked. For example, if the expression is high in two conditions A and B, then the ratio the expression A/B might be small although the difference $A-B$ can be large. In contrast, if the expression of A and B is very low, even a small difference $A-B$ might result in a large ratio A/B. Therefore, fold changes are usually complemented by measures of effect significance.

Statistical tests for differential expression

The significance of an effect is a common criterion for ranking and filtering genes in microarray experiments. Statistical tests calculate p-values, which represent the probability of seeing an equal or higher expression difference between a group of samples and a group of controls under a null hypothesis. The choice of the null hypothesis depends on the applied test and may differ for different software and algorithms. Measured expression levels are often non-normally distributed, and their distributions may be dependent and non-identical between genes (Smyth, 2004). Furthermore, the large number

of genes on a typical microarray presents a multiple testing problem, which requires the consideration and application of multiple testing correction methods.

To assess effect significance in microarray experiments, Smyth (2004) propose the use of empirical Bayes methods, which can borrow information from the ensemble of genes to infer the variance of measurements of individual genes. Smyth (2004) also reformulate the posterior odds statistic that is used by the Bayes method into a moderated t-statistic, which allows the use of posterior residual standard deviations instead of ordinary standard deviations and requires less hyperparameters to be estimated. By combining the Bayesian approach with the t-statistic and providing closed form equations to estimate the hyperparameters, Smyth (2004) extend an approach that was first proposed by Lönnstedt and Speed (2001) to arbitrary experimental designs and microarray types. In short, Smyth (2004) assume a linear model $\mathrm{E}[\mathbf{y}_j] = \mathbf{X}\alpha_j$, where $\mathbf{y}_j$ is the expression data for gene j, $\mathbf{X}$ is the design matrix, and α_j is the vector of coefficients. The contrasts, i.e., the sample groups that are compared, are defined by $\beta_j = \mathbf{C}^T\alpha_j$, where $\mathbf{C}$ is the matrix of contrasts. By fitting the linear model $\mathrm{E}[\mathbf{y}_j]$, the coefficients α_j are estimated, which, in turn, enables the estimation of β_j. The full details of the p-value calculation are provided by Smyth (2004).

In this thesis, we used the `limma` package for R/Bioconductor (Smyth, 2005) to calculate p-values and corrected for multiple testing with the method proposed by Benjamini and Hochberg (1995). As mentioned above, filtering genes based solely on significance may lead to non-reproducible gene lists across experiments, whereas the fold change ranking offers better reproducibility (Shi *et al.*, 2006). For this reason, we used gene fold changes as our primary criteria for selecting deregulated genes and incorporated p-values where possible. In particular, for our predictive toxicogenomics studies, we rely mainly on the fold change to build machine learning models that predict adverse effects. The following section provides an overview of the machine learning methods that we employed for model building and the validation strategies with which we assessed model performance.

3.2 Machine learning methods for classification

Machine learning is a branch of computer science which encompasses learning algorithms and artificial intelligence. During the learning phase, mathematical and heuristic approaches are used to extract rules and patterns from the training data to construct predictive, data-driven models. In contrast to static, programmed predictors, the prediction is defined not only by the new data but depends on the training data that was used in the model construction process. A large number of algorithms have been explored in machine learning, inspired by optimization theory (support vector machines, Boser *et al.* (1992)), data structures (decision trees), or biology (artificial neural networks, McCulloch and Pitts (1943)). Machine learning methods employ various techniques from many other branches of computer science and statistics and have, vice versa, been applied to

many tasks in other fields such as spam detection (Laskov and Šrndić, 2011) or computer vision (Jiang *et al.*, 2016). A distinction is made between supervised and unsupervised machine learning. In supervised learning, labels are provided for each training instance, whereas the labels are unknown in unsupervised learning. Thus, supervised learning algorithms are prediction or regression methods that provide an output for new instances of data. In contrast, unsupervised learning, also called data mining, is used for exploratory data analysis and searches for hidden structure in the data.

Predictive toxicogenomics employs both supervised and unsupervised learning for different types of studies. The studies presented in this thesis built predictive models with different, supervised machine learning algorithms, which are described in the following sections.

3.2.1 Support vector machines

Support vector machines (SVMs) are among the most popular machine learning algorithms and have been applied in many application areas. For example, SVMs are used in bioinformatics for protein structure prediction (Cai *et al.*, 2002) or prediction of adverse drug effects (Ellinger-Ziegelbauer *et al.*, 2008). Vapnik and Lerner (1963) developed the first version of an SVM and Cortes and Vapnik (1995) published the soft margin SVM, which is most frequently used today. The following description is based on a review of SVMs in computational biology by Ben-Hur *et al.* (2008).

SVMs are maximum-margin classifiers, i.e., they maximize the distance of all points in each class to a hyperplane that separates the classes (Fig. 3.2). The hyperplane is the decision boundary for classification and depends only on a subset of the training vectors, which are called support vectors and lie on the margin. One major constraint on the hyperplane is linearity in the input feature space. For this reason, the hyperplane can only separate classes if the underlying decision problem is linearly separable, which for many relevant applications is not the case. To overcome this limitation, the input features can be transformed into a higher-dimensional space in which the data becomes linearly separable (Boser *et al.*, 1992). This transformation is often referred to as the "kernel trick" and provides an elegant solution for applying SVMs to non-linear problems. However, the transformation into higher-dimensional kernel spaces increases the generalization error and thus requires more training examples. Due to the limited number of training samples that is available in toxicogenomics studies, the experiments in this thesis used only linear SVMs.

Linear SVMs learn hyperplanes in the input feature space and use a linear discriminant function $f(\mathbf{x})$, which is defined as

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b, \tag{3.1}$$

where $\mathbf{x}$ are data vectors, $\mathbf{w}$ is the weight vector, b is the bias, and $\langle \cdot, \cdot \rangle$ is the dot product. The points $\mathbf{x}$ that satisfy $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$ define the hyperplane. For each point $\mathbf{x}$, the

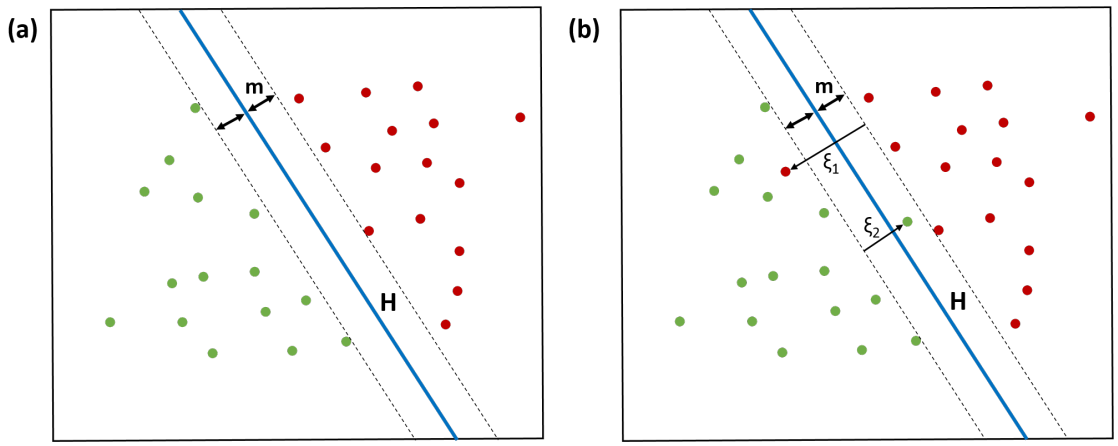


Figure 3.2: **Linear support vector machine.** Linear SVMs separate classes with a linear hyperplane **H** that maximizes the margin **m** between classes. The hyperplane is defined by the equation $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$, where $\mathbf{w}$ are the learned weights. New observations are classified based on their position to the hyperplane. The hard-margin SVM (a) can only learn linearly separable problems. By introducing slack variables ξ, the soft-margin SVM (b) can also learn problems that are not linearly separable.

corresponding label y is the sign of the discriminant function $f(\mathbf{x})$. By convention the labels y are 1 for the positive class and -1 for the negative class. The training problem for an SVM is to find the weight vector $\mathbf{w}$ and bias b which maximize the margin and minimize the training error given a set of n training instances $\mathbf{x}_i$ with associated labels y_i for $i \in \{1, \ldots, n\}$. The margin is the smallest distance between any point $\mathbf{x}_i$ and the hyperplane defined by $f(\mathbf{x}) = 0$, and the training error is the number of points for which $\text{sign}(f(\mathbf{x}_i)) \neq y_i$, i.e., which are located on the wrong side of the hyperplane.

A classification problem is called linearly separable if a hyperplane exists such that $\text{sign}(f(\mathbf{x}_i)) = y_i$ for all instances $\mathbf{x}_i$. For linearly separable classification problems, the hard-margin SVM solves the training problem by maximizing the margin, which provides the best generalization that is possible based on the training instances (Fig. 3.2(a)). The optimization problem for the hard margin SVM can be solved with solvers based on convex optimization theory and is given by

$$\begin{aligned} &\min_{\mathbf{w},b} \frac{1}{2} ||\mathbf{w}||^2 \\ \text{s.t.} \quad &y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b \right) \geq 1, \text{ for } i = 1, \ldots, n \end{aligned} \tag{3.2}$$

While the hard-margin SVM can be efficiently applied for linearly separable problems, most real-world classification problems are not linearly separable. This can be due to measurement noise, problem complexity, or other reasons. To overcome this limitation, Cortes and Vapnik (1995) developed the soft-margin SVM, which allows classification

errors during the training (Fig. 3.2(b)). The soft-margin SVM introduces slack variables ξ_i, which leads to an optimization problem similar to the hard-margin SVM:

$$\begin{aligned} &\min_{\mathbf{w},b,\xi} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad &y_i\left(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\right) \geq 1 - \xi_i,\ i = 1,\ldots,n \\ &\xi_i \geq 0,\ i = 1,\ldots,n \end{aligned} \tag{3.3}$$

The introduction of the slack variables in the constraints relaxes the hard constraints of the hard-margin SVM. To minimize the classification error, the term $C\sum_{i=1}^{n}\xi_i$ in the objective function penalizes classification errors. The sensitivity to training errors can be adjusted by setting the parameter C, which represents the trade-off between fitting the maximum-margin hyperplane and the generalization of the resulting discriminant function. The parameter C is problem-specific and is often selected by parameter optimization.

To solve the optimization problem in Eq. (3.3), the problem is transformed into the dual problem, which is a maximization problem:

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j\rangle \\ \text{s.t.} \quad &\sum_{i=1}^{n} y_i \alpha_i = 0, \text{ with } 0 \leq \alpha_i \leq C \end{aligned} \tag{3.4}$$

This maximization problem can be solved efficiently with gradient descent methods. With the optimal solution for α and the training instances, the weight vector $\mathbf{w}$ can be calculated:

$$\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i \tag{3.5}$$

In the solution of the dual problem, $\alpha_i \neq 0$ only for support vectors. The remaining training instances can be omitted after the learning phase, which provides a sparse representation of the dataset and enables a more efficient classification. Finally, new instances $\mathbf{x}$ are classified by calculating $\text{sign}(f(\mathbf{x}))$ as defined by Eq. (3.1).

3.2.2 Neural networks

Artificial neural networks are a group of machine learning algorithms that were initially inspired by the architecture of the human brain. We used multi-layer perceptrons (MLPs), feed-forward neural networks that extend on the idea of a perceptron. This description of MLPs is based on the book *Pattern Recognition and Neural Networks* by

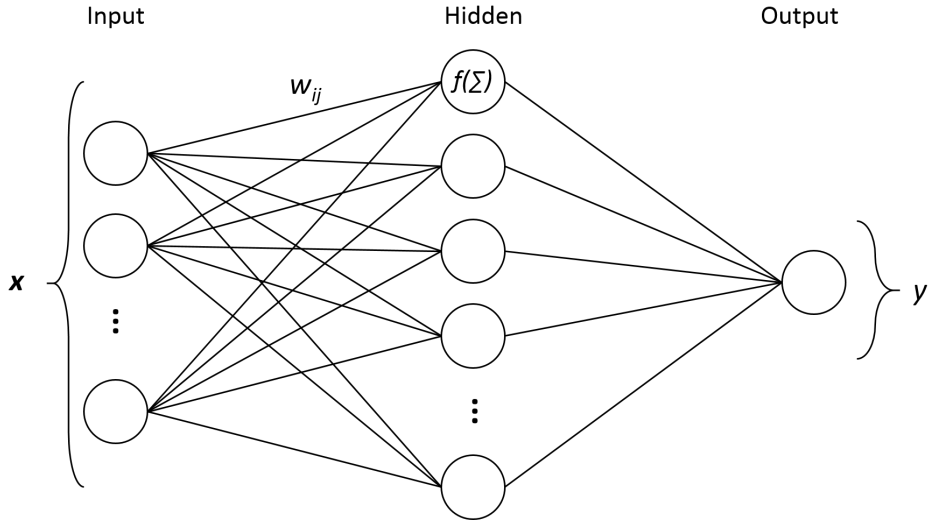


Figure 3.3: **Multilayer perceptron (MLP) with one hidden layer.** An MLP consists of several layers of neurons that are fully connected. The connection between neurons i and j are weighted by a weight w_{ij}. Each neuron sums over all inputs and applies a function f_j to calculate the output which is propagated to the next layer. The features **x** are presented to the input layer and the predicted class y is observed at the output layer.

Ripley (1996). Feed-forward networks consist of numbered units, which are connected with other units by one-way links such that each unit is connected only to units with higher numbers. For these networks, the units can be arranged in layers such that only connections between layers exist (see Fig. 3.3), which also provides the name MLP. Each unit sums all inputs received from incoming connections into a value x_j and applies a function f_j to determine the output y_j, which is fed to the following units. All connections are weighted by a factor w_{ij}. Traditionally, MLPs were often used with three layers: the input layer that is determined by the feature representation, a single hidden layer, and the output layer. MLPs can be trained with backpropagation, which uses the difference between the observed output and the known, desired output to modify the connection weights and minimize the training error. Recently, deep neural networks, which use many hidden layers, achieved good results on many classification tasks. However, deep neural networks require a large number of training examples. Due to the limited number of samples in toxicogenomics studies, we used MLPs with only one hidden layer as implemented in the `nnet` package for R (Venables and Ripley, 2002).

3.2.3 Random forests

Random forests (RFs) comprise a group of classification methods which use multiple independent decision trees to make predictions. Classification methods such as RFs are also called ensemble methods because they employ an ensemble of smaller classifiers

that contribute to the classification. The name "random forest" was first used by Breiman (2001), who provides the following definition for a RF: "A random forest is a classifier consisting of a collection of tree-structured classifiers $\{h(\mathbf{x}, \Theta_k), k = 1, \ldots\}$ where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input $\mathbf{x}$". Breiman (2001) demonstrates that the generalization error converges as the number of trees in the RF increases, which indicates that RFs do not overfit as more trees are added. RFs are parametrized by the number of decision trees n_{tree} and the number of features m_{try} which are used at internal splits in the decision trees. However, results by Breiman (2001) show that RFs are not very sensitive to the parameter values and good, robust classification results are obtained with standard parameter values. A general algorithm for building a RF is provided by Liaw and Wiener (2002) as follows: First, draw n_{tree} bootstrap samples from the original data. Then, for each bootstrap sample, grow an unpruned tree, where at each node m_{try} of the features are chosen to compute the best split. Finally, the new data is predicted by collecting the votes on all n_{tree} trees. For each bootstrap sample, a random subset of training instances is chosen to ensure the independence of the trees in the forest. The out-of-bag samples, i.e., the training instances that are not selected for the tree building, are used to provide an estimate of the generalization error. Implementations of RFs vary in the method for selecting the features at each internal node, the proportion of features and instances in the bootstrap sample, and the underlying decision tree algorithm. Nevertheless, the results on convergence and overfitting by Breiman (2001) and the general algorithm by Liaw and Wiener (2002) are shared by most implementations.

3.2.4 Validation of classification models

The generalization error is the most important measure for the evaluation and comparison of classification models. To estimate the generalization error, a classification model is applied to problem instances that have not been used in the classifier training. These left-out instances are referred to as test or validation samples. Test samples need to be drawn from the same distribution as the training samples and must be labeled with the same classes and criteria. For this reason, the training and validation samples are selected from a single dataset in most cases. To ensure that the estimated classification error provides an unbiased estimate, training and validation must be independent, and each sample must only be used in either training or validation. If a dataset contains samples that are not independent (e.g., technical or biological replicates), these need special consideration and have to be grouped to ensure that all replicates are in either training or validation set.

The following sections describe and compare the three most frequently used validation strategies: the holdout method, bootstrapping, and cross-validation (CV). All descriptions are based on the analysis by Kohavi (1995).

Holdout validation

Holdout validation divides the dataset into a training and a validation set. The classification model is build using only the training set. After the training, the model is applied to the validation set to obtain class predictions which are compared to the actual classes to estimate the classification error. However, Cawley and Talbot (2010) show that sampling a single partition of data can arbitrarily favor a particular classifier, which is called sampling bias. Also, Kohavi (1995) argues that the holdout method is a pessimistic estimator because classification power is expected to increase with the number of available training samples. Both problems are particularly problematic for the small datasets that are used in toxicogenomics studies. Also, the holdout method does not provide an estimate of the variance of classifier performance, which is a critical factor for comparing classifiers (Cawley and Talbot, 2010). For these reasons, resampling methods are used to estimate the variance of the classification error and avoid sampling bias.

Bootstrapping and cross-validation

A bootstrap sample of a dataset with n samples is obtained by sampling n samples with replacement. Assuming a uniform sampling, the expected number of distinct samples in a bootstrap is $0.632n$ (Kohavi, 1995). The selected samples are used as the training set, and the remaining samples are used as the test set. To estimate the average performance and the variance, the bootstrapping procedure is repeated.

In k-fold CV, a dataset is divided into k equal sized, pair-wise distinct partitions (called folds) Each fold is used as validation set while the classifier is trained on the remaining $k-1$ folds. Thus, in a k-fold CV, k independent estimates of the classification error are obtained, which provides an estimate of the performance and variance. Stratified CV is an extension of regular CV that ensures that the class distribution is approximately the same in all folds. A special case is n-fold CV or leave-one-out validation, where each sample is held out once and the classifier is trained on all remaining samples. However, Kohavi (1995) argues that replicated k-fold CV with different folds produces more robust estimates of the variance. He also shows that the CV of the estimate is unbiased for a stable learning algorithm. A learning algorithm is stable if the trained model produces the same predictions independently of the permutation of the training data. Furthermore, Kohavi (1995) observed that the variance of k-fold CV does not depend on k.

For a set of real-world datasets with different characteristics, Kohavi (1995) demonstrates that k-fold CV with $10 \leq k \leq 20$ reduces the variance but increases the bias, whereas smaller $k \leq 5$ increase variance due to the instability of the training set. He also concludes that stratified CV is superior to regular CV and recommends replication of the CV evaluation. Bootstrapping has low variance but he observed a large bias in the performance estimates for some of the problems. Based on these findings, he recommends stratified 10-fold CV for model selection.

3.3 Toxicogenomics for preclinical risk assessment

Several studies have evaluated the performance of *in vivo* short-term rodent bioassays for prediction of carcinogenic potential. These studies used microarrays to measure the transcriptional response to the chronic or acute administration of compounds with well-defined carcinogenic potential. Figure 3.4 shows the standard workflow of toxicogenomics studies. The time point of transcriptomic profiling varied among the studies, ranging from 24 hours to 13 weeks. Most predictive toxicogenomics studies used male rats as the model organism and the liver as the target organ for tumor induction. For this reason, most available data was collected from the liver of male rats. Nevertheless, female rats and other organs have been explored in toxicogenomics studies, as well as the mouse as a possible model organism. *In vitro* studies have also been performed by some research groups. The following section reviews the currently available toxicogenomics literature for rat, mouse, and *in vitro* studies.

3.3.1 *In vivo* rat studies

Nie *et al.* (2006) performed the first large predictive toxicogenomics study. They administered a single dose of 138 compounds to male Sprague-Dawley rats. Of the 138 compounds, 24 NGCs and 28 NCs were selected for model building. Nie *et al.* (2006) used statistical and heuristic feature selection followed by an exhaustive search to identify six gene markers. They estimated an accuracy (ACC) of 88.5% with 10-fold CV.

Fielden *et al.* (2007) treated male Sprague-Dawley rats with 147 compounds for up to 7 days. They used 25 NGCs and 75 NCs to build a linear classification model with 31 marker genes. Validation with 20 random bootstraps (60/40 split) resulted in 84.5% ACC, with 56% sensitivity and 94% specificity. The external validation with the remaining 47 compounds resulted in 78.7% ACC (71.4% sensitivity and 84.6% specificity).

The studies by Nie *et al.* (2006) and Fielden *et al.* (2007) were subsequently externally validated in a collaborative study by Fielden *et al.* (2008). The signature proposed by Nie *et al.* (2006) achieved an ACC of 63.9% on the Fielden *et al.* (2007) dataset and 55.1% on an additional proprietary database. The signature proposed by Fielden *et al.* (2007) achieved an ACC of 71.9% on the Nie *et al.* (2006) dataset and 63% on the proprietary database. Based on these findings, Fielden *et al.* (2008) argued that the performance was insufficient to advocate routine use of the signatures.

Ellinger-Ziegelbauer *et al.* (2008) reported a toxicogenomics study with male Wistar rats treated daily with 29 compounds (9 GCs, 11 NGCs, and 9 NCs) for up to 14 days. They used 13 compounds (5 GCs, 5 NGCs, and 3 NCs) to extract signatures and build SVM models. For the best signature, they reported an ACC of 88% for the classification of the validation compounds.

Uehara *et al.* (2008, 2011) performed two predictive toxicogenomics studies with data from the Toxicogenomics Project-Genomics Assisted Toxicity Evaluation system database (TG-GATEs, Uehara *et al.* (2010)). TG-GATEs encompasses gene expression

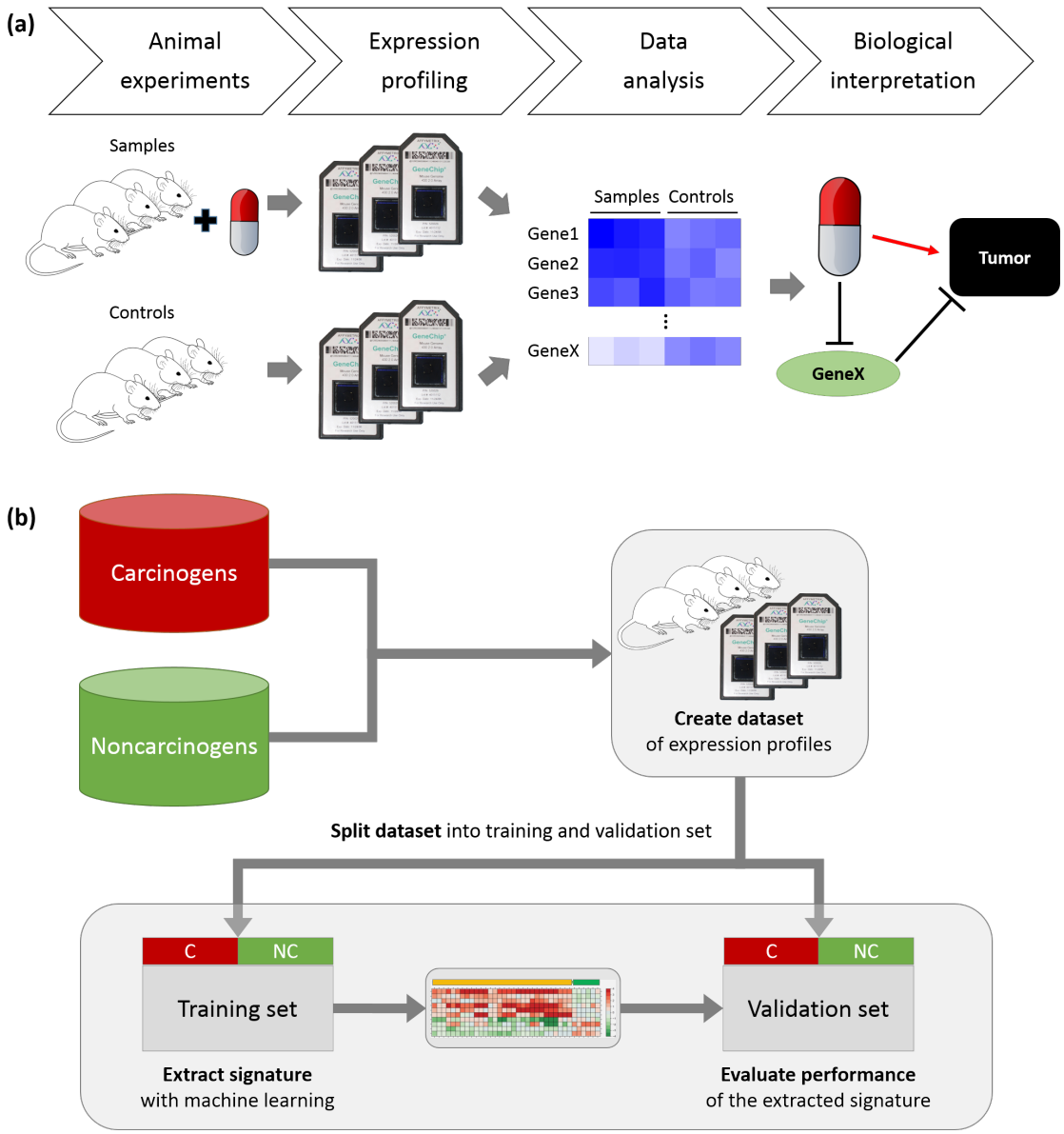


Figure 3.4: **Workflow in toxicogenomics studies.** (a) Expression profiling workflow. Substances are subchronically administered to rats or mice. After the treatment phase, animals are sacrificed and genetic material is extracted from the target tissue. Gene expression is quantified with high-throughput expression profiling methods. The expression profiles in treated and untreated animals are compared and filtered statistically. Relevant expression changes are used for mechanistic analysis. (b) Predictive toxicogenomics workflow. A dataset of expression profiles is collected for carcinogenic and non-carcinogenic substances. Next, the dataset is divided into training and validation set and machine learning or statistical methods are used to extract a biomarker signature from the training set. Then, the predictive performance is evaluated with the validation set.

profiles from male Sprague-Dawley rats treated with 150 compounds. First, Uehara *et al.* (2008) used two NGCs and six NCs for prediction analysis for microarrays (PAM). An external validation of the model with 22 additional compounds (8 NGCs and 14 NCs) resulted in 73.3% ACC (37.5% sensitivity, 95% specificity). In 2011, Uehara *et al.* (2011) reported a new model that was built using a larger subset of the TG-GATEs data. Using 60 compounds (6 NGCs and 54 NCs), they used statistical feature selection and trained an SVM model. The external validation on the remaining TG-GATEs compounds yielded an ACC of 78.9% (15.8% sensitivity and 100% specificity). Further, Uehara *et al.* (2011) applied their model to 14 compounds from the NEDO project (Matsumoto *et al.*, 2009), which resulted in 50% ACC (30% sensitivity, and 100% specificity).

The NEDO project (Matsumoto *et al.*, 2009) treated male F344 rats with 88 compounds for up to 4 weeks. In 2011, Hiroshi Matsumoto (2011) reported an SVM-based model trained on a subset of 41 compounds (17 Cs, 24 NCs). They performed external validation on the 45 compounds not used for training, which resulted in an ACC of 28.8% (11.8% sensitivity and 81.8% specificity). Following the initial study, Matsumoto *et al.* (2014) reported a new SVM model, which extended the number of compounds used for training to 68 (46 Cs and 22 NCs). The new model achieved an ACC of 68.8% on the validation data (72.7% sensitivity and 60.0% specificity).

In 2010, Auerbach *et al.* (2010) argued that carcinogenesis is a transient process, which may not be detected after few weeks of drug administration. Using male F344 rats, they sampled gene expression profiles after up to 90 days of daily exposure to 10 compounds (4 Cs and 6 NCs). They used SVMs and recursive feature elimination (RFE) and reported 100% ACC after leave-one-out validation for most models. However, external validation on an independent dataset resulted in accuracies ranging from 74.1%-83.8% when using all time points and 84.5% to 90% when using data obtained after 90 days.

Following a previous study with *in vivo* mouse data, Eichner *et al.* (2014a) evaluated ensemble feature selection (EFS) on the 14-day data of the TG-GATEs dataset. They used 20 compounds (9 NGCs and 11 NCs) for training and applied linear SVMs, PAM, RFE, and Golub's signal-to-noise ratio to extract signatures. With three-fold CV, they estimated a mean area under the receiver operating characteristic (AUC) of 0.95 for their signature. They also applied EFS to four additional datasets and demonstrated improved prediction ACCs compared to published signatures.

Recently, Gusenleitner *et al.* (2014) used TG-GATEs and DrugMatrix data to build RFs with 500 genes selected by a variance filter. Using a 70/30 bootstrap split approach with 200 resamplings, they estimated an AUC of 0.77 on the DrugMatrix dataset and 0.83 on the TG-GATEs dataset. A prediction of the TG-GATEs data with the model built on the DrugMatrix dataset yielded an AUC of 0.77, confirming the robustness of the model and its estimated performance. Additionally, Gusenleitner *et al.* (2014) explored the use of perturbed pathways as potential markers, which resulted in a slightly lower AUC in the internal performance evaluation (AUCs of 0.73 for DrugMatrix and 0.81 for TG-GATEs), but better performance in the cross-data-set evaluation (AUC of 0.79).

3.3.2 *In vivo* mouse studies

Several toxicogenomics studies have also investigated the use of the mouse as a model organism. Thomas *et al.* (2007) profiled lung tissue of female B6C3F1 mice exposed to 7 lung carcinogens and 6 noncarcinogens. They used 10-fold CV to estimate an ACC of 93.9% (95.2% sensitivity and 91.8% specificity). However, these results are likely positively biased due to the presence of biological replicates in both training and test sets. In 2009, Thomas *et al.* (2009) reported another study with 15 lung carcinogens and 11 noncarcinogens. They built a total of 84 different models, with the best achieving 77.5% ACC, estimated by an unbiased five-fold CV.

In 2014, Melis *et al.* (2014) profiled the liver of male C57BL mice after 28 days of repeated administration of seven GCs, nine NGCs, and eight NCs. They build a multi-class prediction model from a subset of 16 compounds (4 GCs, 7 NGCs, and 5 NCs). External validation of the model on the remaining 8 compounds resulted in 50% ACC (80% sensitivity and 0% specificity).

Eichner *et al.* (2013a) treated male and female CD-1 mice with three GCs, three NGCs, four NCs and used EFS to extract marker genes. Evaluation with a bootstrapping approach yielded a mean AUC of 0.92 for C vs. NC classification and 0.83 for GC vs. NGC classification.

3.3.3 *In vitro* studies

In vitro bioassays have also been proposed to replace animal experiments entirely. However, *in vivo* models remain the preferred method for chronic exposure tests because the metabolic processing of drugs is reduced in *in vitro* systems (Waters *et al.*, 2010). Nevertheless, *in vitro* assays provide a rapid and resource-efficient strategy for risk assessment and can be applied to human cells to reduce problems introduced by cross-species differences. *In vitro* toxicogenomics studies have used both human and rodent cells.

Among the first to explore the use of *in vitro* assays for predictive toxicogenomics were van Delft *et al.* (2005). They treated human HepG2 cells with nine genotoxins and seven nongenotoxins for 24 hours and trained four classifiers. The ACC ranged from 33% to 83% for in the external validation with six compounds.

In 2011, Yildirimman *et al.* (2011) treated hepatocyte-like cells with 15 compounds (5 GCs, 5 NGCs, and 5 NCs) and derived a 592-gene signature. Using leave-one-out, Yildirimman *et al.* (2011) reported an ACC of 93.8% for their model. However, due to the inclusion of up to four replicates for each compound, this result is likely positively biased due to the presence of replicates in the training set.

In a follow-up study, Doktorova *et al.* (2013) used five different cell types: human embryonic stem cells, HepG2 cells, HepaRG cells, and two rat hepatocyte cell types. They observed good correspondence of gene expression changes with carcinogenicity but an identification of NGCs was not possible.

Magkoufopoulou *et al.* (2012) used human HepG2 cells to predict *in vivo* genotoxicity. They treated HepG2 cells with 62 compounds for up to 48 hours and used 34 compounds for model training with PAM. When using only transcriptomics data, they achieved ACCs of 79% (leave-one-out on training set) and 57% (external validation).

In a series of studies, Schaap *et al.* (2014, 2012) explored the use of *in vitro* mouse models. They treated murine cells with 26 compounds (16 NGCs and 10 GCs/NCs) for up to 72 hours. Schaap *et al.* (2014, 2012) focused on unsupervised analysis and MOA detection by defining pairs of NGCs with similar MOAs and using T-statistics to recall the matching partner for each compound but did not report performance measures.

3.3.4 Limitations of toxicogenomics studies

As described above, many studies have demonstrated that predictive toxicogenomics can be used to assess the carcinogenic potential of compounds. Nevertheless, some limitations have to be taken into account when interpreting the reported results and considering predictive toxicogenomics for integration into preclinical risk assessment. In their review, Waters *et al.* (2010) highlighted several problems concerning the compound annotation. Genotoxicity is commonly defined as Ames positive, whereas Ames negative compounds are considered nongenotoxic (Jackson *et al.*, 1993). However, some Ames negative compounds have been shown to induce mutations in mammalian cells and chromosomal damage, cell transformation, or loss of alleles in other *in vivo* or *in vitro* genotoxicity assays. Waters *et al.* (2010) refer to these compounds as Ames negative genotoxic carcinogens. Currently, it has not been concisely evaluated if this should be considered in the labeling of compounds for training sets and might have caused confounding effects in previous studies that aimed at distinguishing genotoxic and nongenotoxic compounds.

Also, the relevance of the LRB for human risk assessment has been called into question (Cohen, 2010; Ward, 2007). Knight *et al.* (2006a) performed an extensive evaluation of 160 compounds for which carcinogenicity testing had been conducted in animals. Mice (92.4%) and rats (86.7%) were used in most long-term bioassays, and the liver (66.3%) was the organ affected by chemicals most often. However, upon inspection of chemicals found to be carcinogenic in rodents that have also been assessed in monkeys, only half were found to be carcinogenic in monkeys. Monro and Mordenti (1995) list substantial differences between humans and rodents such as mean lifespan, food consumption, basal metabolic rate, anatomical differences, and DNA excision repair rate that might confound the prediction of human risk from rodent tumor findings. An analysis of 61 chemicals that are carcinogenic in rats or mice showed that only 13 chemicals were carcinogenic in both species and Knight *et al.* (2006a) conclude that "the profound discordance of bioassay results between rodent species, strains, and genders, and further, between rodents and human beings, means that it is profoundly difficult to make human carcinogenicity assessments on the basis of rodent bioassay data." (Knight *et al.*, 2006a). Nevertheless, the LRB is still used by pharmaceutical companies and required by regulatory agencies

for drugs with prolonged or chronic administration. Knight *et al.* (2006b) also propose alternatives for the LRB. Among others they also put forward toxicogenomics, which offers more information on MOAs than raw tumor incidence in the bioassay. Nevertheless, the question of how to translate positive findings from future toxicogenomics based assays to human risk assessment is still discussed by experts (Ellinger-Ziegelbauer *et al.*, 2009).

The previous limitations apply to toxicogenomics in general and are being assessed by experts in the field. However, there are also problems that are specific to studies in predictive toxicogenomics. These problems concern the evaluation of proposed models. First, the sample size is very low in many studies due to the resources (e.g., money, time, animals) necessary to generate the gene expression profiles for toxicogenomics studies. The first toxicogenomics studies sampled data for as few as six (Iida *et al.*, 2005) or nine compounds (Kramer *et al.*, 2004). Meanwhile, larger databases have been released: TG-GATEs with 150 compounds (Uehara *et al.*, 2010), DrugMatrix with 130 compounds (Ganter *et al.*, 2005), the dataset by Nie *et al.* (2006) with 129 compounds, and the NEDO dataset with 88 compounds (Matsumoto *et al.*, 2009). However, many recent studies still collect expression profiles for less than 30 compounds (e.g., Thomas *et al.*, 2011; Eichner *et al.*, 2013a; Ellinger-Ziegelbauer *et al.*, 2008; Melis *et al.*, 2014) or use only a small subset of larger databases for training (e.g., Uehara *et al.*, 2011, 2008). In addition, NGCs act through a wide range of mechanisms (Silva Lima and Van der Laan, 2000). Thus, many models built with a low number of NGCs as positive training compounds are likely to suffer from low sensitivity. Notable examples are the studies by Uehara *et al.* (2008, 2011), which are performed with only two and six NGCs with very similar MOAs as training compounds and consequently yield low sensitivities. To improve the sensitivity, more NGCs must be included in model generation and biomarker extraction. Also, the training compounds should cover a broad range of MOAs to ensure generalization.

Second, many toxicogenomics studies focus on a single dataset and do not perform extensive validation of biomarkers, although large toxicogenomics databases are available (Ellinger-Ziegelbauer *et al.*, 2008; Hiroshi Matsumoto, 2011; Matsumoto *et al.*, 2014; Schaap *et al.*, 2014). Possible reasons are the choice of different microarrays, time points, doses, or animal strains. However, these problems can be addressed by mapping probes between microarrays or choosing the most similar time points and doses. This was demonstrated early in an extensive external validation study performed by Fielden *et al.* (2008), who found that the reported model performances within the same dataset were higher than those observed in the inter-laboratory evaluation. This early finding should stimulate more caution when reporting and interpreting the performance of models and biomarkers that have only been evaluated internally.

Third, the reported model accuracies are biased in some predictive toxicogenomics studies due to flaws in the evaluation scheme. We identified three main sources of bias: (1) the presence of samples from the same compound in training and test set (e.g., at different time points or doses) (e.g., Uehara *et al.*, 2011, 2008; van Delft *et al.*, 2005;

Yildirimman *et al.*, 2011), (2) feature selection based on the whole dataset and subsequent CV of the selected markers (e.g., Jonker *et al.*, 2009; Yildirimman *et al.*, 2011; Hiroshi Matsumoto, 2011; Matsumoto *et al.*, 2014), which introduces selection bias (Cawley and Talbot, 2010), and (3) exclusion (e.g., Fielden *et al.*, 2007; Uehara *et al.*, 2011; van Delft *et al.*, 2005) and mislabeling (e.g., Uehara *et al.*, 2011) of compounds. The mentioned flaws do not necessarily lead to positive bias, but can inflate the estimated performance, mask over-fitting, and reduce the usefulness of the model for risk assessment of unseen compounds. For this reason, predictive toxicogenomics studies that use internal validation should use CV or bootstrapping where the whole process of model building (signature extraction and classifier training) is performed anew for each data split (Cawley and Talbot, 2010). Also, if multiple time points, doses, or replicates are available for a compound, all should be in either the training or the test set, not both. Further consideration is necessary when evaluating the model performance in small studies (less than 20 compounds), as the choice of the evaluation model (e.g., 10-fold CV or leave-one-out) can significantly influence the estimated performance (Airola *et al.*, 2009).

Chapter 4

The ZBIT Bioinformatics Toolbox for computational biology

During the last 20 years, bioinformatics has become ever more present in biology and biomedical research. At first, bioinformatics was required most prominently in genomics, in particular for sequence analysis. In sequence analysis, bioinformatics not only enabled the reconstruction of the human genome from shotgun sequencing, but also enabled the computational analysis of proteins or proteomics, e.g., for transcription factor identification (Wasserman and Sandelin, 2004). However, as the processing power of computers increased and new high-throughput methods were established, bioinformatics gained access to many branches of biology. With the development of cDNA microarrays in the 1990s, gene expression analysis (or transcriptomics) was added to the repertoire of genomics studies (Schena *et al.*, 1995). The genome-wide gene expression screening stimulated bioinformatics research in unsupervised and supervised pattern analysis (Brazma and Vilo, 2000). More recently, reverse phase protein arrays (RPPAs) enabled the profiling of protein abundance and post-translational modifications (PTMs) (Spurrier *et al.*, 2008). The improving methodology and accuracy along with decreasing prices and ever more powerful computation clusters prompted research into the simulation of life. Systems biology has emerged as the primary branch for developing models of living organisms on various scales (Kitano, 2002). The ultimate goal of systems biology is the development of predictive, *in silico* models of living organisms, which could be used to identify causes of metabolic diseases, develop target-specific drugs, or assess the potential side effects of drugs *in silico*. Recently, Karr *et al.* (2012) reported a whole-cell computational model to prove the viability of this goal. These simulation experiments usually involve measurements of hundreds of metabolites to determine the internal model structure and constraints. Overall, the advent of *omics* technologies and big data led to great breakthroughs but also created new challenges. As high-throughput technologies provided data at unprecedented scale and detail, new methods for automated analysis and dimensionality reduction were necessary. This led to the development of new analysis methods and computational frameworks designed to process big data.

Whereas the technology made great leaps in these 20 years, bioinformatics software has retained a high entry burden for interested life science researchers. Commercial

bioinformatics software is available for many standard analysis tasks, particularly for gene expression analysis or genomics. Nevertheless, most bioinformatics tools are academic software developed for very specific tasks and often have a clearly defined user group: bioinformaticians. For this reason, advanced technical knowledge is often required to efficiently use bioinformatics tools. For example, many tools depend on particular operating systems or require third-party libraries that are not included with the software and have to be installed separately. Some tools are only available as source code and have to be compiled by the user. Many tools define custom file formats with more or less strict specifications, which often means that a parser or converter must be written to use a tool or its output. In addition, graphical user interfaces are often not provided, such that knowledge of how to use the command line is required. Documentation and usage examples may also be missing or be reduced to a bare minimum. This also translates to computational frameworks for big data, which require tailor-made software that can use the framework, technical knowledge for setup and maintenance, and the necessary infrastructure such as computation clusters and database servers.

Web platforms have been proposed to address the above-mentioned problems and enable life science researchers to use bioinformatics tools and provide bioinformaticians with a simple way to distribute their software. These web platforms provide predefined interfaces to command line tools, which eliminates a number of problems: the user does not need to install the tool or any dependencies locally, he can access the tool through the familiar interface of his web browser, and the actual analysis is performed remotely such that no specific hardware infrastructure is required. This makes the tool execution independent of the user's hardware and allows running computationally expensive analyses remotely from a mobile device in the lab. A prominent example for porting bioinformatics software to an online platform is the public Galaxy server, which is an open, web-based platform for computational biology (Goecks *et al.*, 2010). The public Galaxy server is an instance of the Galaxy framework, which can be used to set up customized web platforms with a different sets of tools. While the framework was initially developed for sequence analysis, it has spread into many other *omics* branches such as proteomics and systems biology (Narang *et al.*, 2014; Hildebrandt *et al.*, 2015). The Galaxy framework provides an established, user-friendly graphical interface for command line tools through the user's browser. For developers, it provides many functions out of the box, e.g., user management, storage of results, and histories of analysis.

With the ZBIT Bioinformatics Toolbox, we have created a customized Galaxy instance that provides a number of bioinformatics tools to life science researchers without strong technical background. This chapter is based mainly on the publication "ZBIT Bioinformatics Toolbox: A Web-Platform for Systems Biology and Expression Data Analysis" in *PLoS ONE* (Römer *et al.*, 2016b).

4.1 Tools included in the ZBIT Bioinformatics Toolbox

The ZBIT Bioinformatics Toolbox provides eight bioinformatics tools, which we have categorized into three major branches of biology: systems biology, transcription factor annotation, and expression data analysis. The tools were originally developed at the chair of Cognitive Systems at the University of Tuebingen for various projects and are available either as Java applications, R scripts, or command line tools. With the ZBIT Bioinformatics Toolbox, we provide a user-friendly interface to these tools for biologists and other life science researchers without a strong technical background. The tools that we included in the ZBIT Bioinformatics Toolbox are BioPAX2SBML, SBMLsqueezer, SBML2LaTeX, and ModelPolisher in systems biology, TFpredict and SABINE for transcription factor analysis, and RPPApipe and ToxDBScan for expression data analysis (see Fig. 4.1).

4.1.1 Systems biology

The Systems Biology Markup Language (SBML, (Hucka *et al.*, 2003)) and the Biological Pathway Exchange format (BioPAX, Demir *et al.* (2010)) are two of the most widely used community standards in systems biology (Dräger and Palsson, 2014). For a long time, both formats were not compatible: while the SBML standard is designed for quantitative analysis, BioPAX is optimized for the exchange of qualitative pathways between databases (Büchel *et al.*, 2012). Researchers proposed and developed many tools to convert models stored in a specific format into another format to facilitate the exchange of models between systems biology groups and databases. In addition, some tools can add information from external databases to improve or extend existing models. The following paragraphs provide short summaries of the systems biology tools that are part of the ZBIT Bioinformatics Toolbox and discuss their capabilities and relevance for researchers.

BioPAX2SBML was the first converter that was able to translate models from BioPAX into SBML and properly conserve the qualitative relations which are defined by BioPAX (Büchel *et al.*, 2012). BioPAX describes the semantics of biological networks and is used mainly for qualitative analysis (Demir *et al.*, 2010). In contrast, SBML describes quantitative models and includes mathematical expressions which are necessary for dynamic simulations (Hucka *et al.*, 2003). Before the definition of the Qualitative Models extension for SBML (`qual`, (Chaouiya *et al.*, 2013)), it was not possible to include qualitative interactions in SBML models. BioPAX2SBML was the first converter that used `qual` to conserve the qualitative information in BioPAX models when converting them to SBML. Büchel *et al.* (2012) implemented the translation by defining a mapping that for each element of the BioPAX model adds the translated SBML element to the converted SBML model. This conversion maintains exact reactions where possible and includes all other interactions as qualitative transitions. By using the KEGG API, the resulting SBML models are further augmented with additional information from external databases such

as KEGG (Kanehisa *et al.*, 2014) and Entrez Gene (Maglott, 2004). Thus, BioPAX2SBML makes the information in curated, qualitative pathway databases accessible for automated processing. For example, the Path2Models project used BioPAX2SBML to create mathematical models in SBML for biochemical pathway maps retrieved from multiple data sources that provide BioPAX models (Büchel *et al.*, 2013).

SBMLsqueezer generates kinetic equations from the stoichiometry, the participating species, and regulatory relations stored in an SBML model (Dräger *et al.*, 2008). These kinetic equations are necessary for the dynamic simulation of models created from qualitative networks with BioPAX2SBML or from graphical representations of models such as the Systems Biology Graphical Notation (SBGN). While the manual creation of models in SBGN is feasible with editors like CellDesigner, the assignment of kinetic rate laws is not straightforward and prone to human errors (Dräger *et al.*, 2008). By automating this step, SBMLsqueezer greatly facilitates model generation. SBMLsqueezer uses the annotations of reactants, products, and regulatory elements in each reaction to generate the correct rate law and supports numerous kinetics. In addition, SBMLsqueezer retrieves experimentally determined rate laws from the SABIO-RK database (Wittig *et al.*, 2012) to extend the available information for simulations. The models augmented with SBMLsqueezer can then be simulated using standard simulation software such as CellDesigner. The Path2Models project used SBMLsqueezer to add kinetic equations to the SBML models created with BioPAX2SBML (Büchel *et al.*, 2013). Other groups have used SBMLsqueezer for various dynamic simulations of biological networks, which can be used to predict the behavior of cells in response to stress factors such as drug administration or heat shocks. Notable examples are Pathak *et al.* (2013), who modeled the MAPK machinery activation in plants, and Gupta and Misra (2013), who simulated the effects of drugs with systems biology approaches.

SBML2LaTeX is an SBML converter that generates human-readable reports for SBML models (Dräger *et al.*, 2009). For this purpose, the XML-based SBML model is translated into a LaTeX document, which can be compiled into either DVI or PDF format for printing or HTML for web pages. SBML2LaTeX facilitates the complicated model development process by providing human-readable reports, which allow easier error-checking than the machine-readable XML files, and facilitating model communication between researchers. The model report can also be included in scientific writing, e.g., as supplemental material to ease model interpretation for external researchers in publications. To translate the SBML model into a PDF report, SBML2LaTeX creates tables of all included reactions, species, constraints, rules, and definitions. These are included in separate sections of the report and linked to relating elements. For example, each reaction is represented by a subsection of the "Reactions" section and lists the involved reactants, the reaction equation, and relevant kinetic laws. The most prominent example for the usage of SBML2LaTeX is the BioModels Database (Li *et al.*, 2010; Chelliah *et al.*, 2015), which uses SBML2LaTeX to automatically generate human-readable PDF reports for each model in the database.

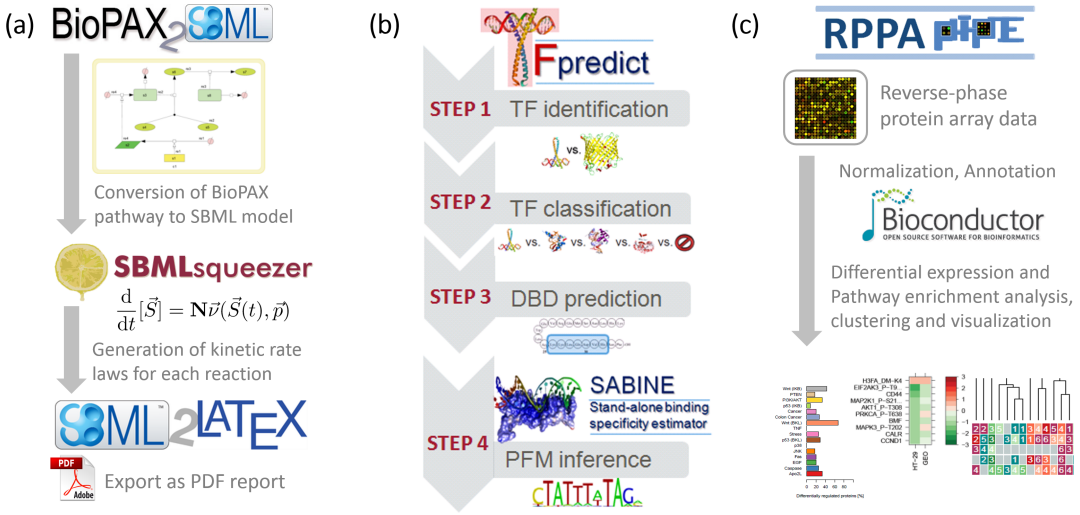


Figure 4.1: **Analysis workflows available in the ZBIT Bioinformatics Toolbox.** The workflows presented in this figure represent fundamental use-case scenarios that combine tools from all three basic classes of tools within the ZBIT Toolbox. (a) SBML model processing with BioPAX2SBML, SBMLsqueezer, and SBML2LaTeX. BioPAX2SBML converts models from BioPAX to SBML and conserves qualitative models. SBMLsqueezer generates kinetic rate laws for each reaction contained in an SBML file. SBML2LaTeX creates human-readable reports from SBML files. (b) Transcription factor analysis using TFpredict and SABINE. TFpredict is used to identify transcription factors and predict their superclass and DNA binding domains. SABINE uses this information to infer the PFM that represents their DNA binding profile. (c) RPPA analysis with RPPApipe. RPPApipe implements a customizable pipeline for RPPA data analysis. This includes normalization and annotation of raw data, statistical methods for the detection of deregulated and differentially modified proteins, and their association with alterations on the pathway level, and visualization of the results. Figure from Römer *et al.* (2016b).

BioPAX2SBML, SBMLsqueezer, and SBML2LaTeX all use the SBML format and are therefore fully compatible. They can be combined into pipelines that create and extend SBML models and provide human-readable summaries for each model (Fig. 4.1a), either from BioPAX models that are converted with BioPAX2SBML or with models that are already available in SBML format.

ModelPolisher is a more recently developed tool that generates well-annotated SBML models based on the BiGG Models knowledge base (King *et al.*, 2016). ModelPolisher enriches SBML models that use the conventions of the constraint-based modeling community with additional annotation by complementing their components with annotations from the BiGG database. MIRIAM annotations and Systems Biology Ontology (SBO) terms (Courtot *et al.*, 2011) are added to individual elements of the model. To this end, ModelPolisher matches the model components against the specification of their BiGG

IDs (see King, 2015) and pulls the available metadata for the corresponding entries in the BiGG database. The BiGG IDs are used to recognize specific reaction and metabolite types for which new SBO terms have been defined, e.g., flux bounds and biomass reactions (King *et al.*, 2016). In addition, ModelPolisher confirms the validity of the SBML syntax and the structural correctness of the SBML model to detect mass balance deficiencies. ModelPolisher exports an updated SBML model that can be used as input for following tools within the toolbox or external tools that support the SBML format.

4.1.2 Transcription factor annotation

Transcription factors (TFs) are proteins that bind the DNA at defined regions, so-called DNA domains, and thereby regulate the transcription of genes. TFs have an important role in gene regulation, and researchers identified TFs that are associated with many diseases, for example, inflammatory lung diseases (Rahman and MacNee, 1998), rheumatic diseases (Firestein and Manning, 1999), or cancer (Darnell, 2002). TFpredict uses machine learning to decide if a protein is a TF or non-TF based on the protein's amino acid sequence (Eichner *et al.*, 2013b). For proteins that are predicted to be TFs, TFpredict also attempts to predict the structural superclass and uses the InterProScan web service to detect the TF's DNA-binding domain (DBD) (Jones *et al.*, 2014). The DBD of a TF is the part of the amino acid sequence that is involved in the DNA binding event during gene expression regulation. The structural superclass is a concept of classification based on the tertiary structure of the DBD of a TF. Generally, four main superclasses are used: Basic Domains, Zinc-coordinating DBDs, Helix-turn-helix, beta-Scaffold Factors (Matys *et al.*, 2006). A fifth superclass comprises TFs that belong in neither of these four superclasses. TFpredict employs a sequence-based machine learning approach to build a prediction model for these characteristics and trains the model on data from the TF databases TRANSFAC (Cartharius *et al.*, 2005) and MatBase (Matys, 2003). Eichner *et al.* (2013b) defined BLAST score percentile features, which are based on a BLAST search for the submitted sequence and a comparison of the results with results obtained for TFs and non-TFs. These BLAST search results are used to build a feature vector for the SVM model that predicts if the submitted amino acid sequence is a TF or not. A similar approach in combination with a multi-class SVM is used to predict which of the five superclasses a predicted TF belongs to. In their publication, Eichner *et al.* (2013b) demonstrated that TFpredict performs better than previously published methods.

SABINE (Stand-Alone BINding specificity Estimator) builds on the results of TFpredict to infer the DNA motif of a TF as a position frequency matrix (PFM) (Eichner *et al.*, 2013b). SABINE predicts the PFM based on the amino acid sequence, the DBDs that were detected with InterProScan, the superclass that was predicted by TFpredict, and the species. SABINE uses support vector regression models to identify other TFs from the training set with well-defined PFMs, which were obtained from TF databases, e.g., TRANSFAC (Cartharius *et al.*, 2005) and MatBase (Matys, 2003). Eichner *et al.* (2013b) determined the similarity of TFs based on evolutionary, structural, and chemical similar-

ities. The PFMs of the best matching TFs are then filtered based on the similarity of their respective PFMs using a dynamic threshold value. The remaining PFMs are then merged with a progressive alignment algorithm by the program STAMP, which constructs a consensus DNA motif. As the last step, SABINE performs an error estimation to assess the confidence of the PFM inference.

In combination, TFpredict and SABINE allow the structural and functional annotation of TFs (see Fig. 4.1b). In wet-lab experiments, NR2C2 and PPARA have been confirmed to be TFs in human hepatocytes, as TFpredict and SABINE had previously predicted (Schröder *et al.*, 2011). Currently, TFpredict and SABINE are limited to eukaryotes, but an extension to prokaryotes is in development.

4.1.3 Expression data analysis

Since their advent in the mid-1990s, DNA microarrays have changed the field of genetics, and high-throughput gene expression analysis is now an integral part of biological research (Brazma and Vilo, 2000; Lenoir and Giannella, 2006). Microarray-like technology is today applied in related areas, such as proteomics, where reverse phase protein arrays (RPPAs) are used in individualized medicine or cancer biology (Gallagher and Espina, 2014; Unterberger *et al.*, 2014).

RPPApipe provides a set of tools for RPPA experiments, which allow preprocessing, annotation, statistical analysis, clustering, pathway analysis, and visualization of RPPA data (see Fig. 4.1c, (Eichner *et al.*, 2014b)). Researchers can easily combine these tools to build workflows that are tailored to their experiments. RPPApipe supports several experimental designs: standard paired condition and control designs as well as more specialized designs with multiple conditions or replicated time-series. The major advantage of RPPApipe in comparison with generic array processing software is the support for RPPA-specific analysis. RPPApipe provides specialized visualizations and analyses, e.g., volcano plots that show the differential modification of proteins or pathway profiles that account for the lower number of analytes on RPPAs compared to transcriptomics studies. RPPApipe offers full compatibility with InCroMAP (Wrzodek *et al.*, 2013), a software for integrated multi-omics visualization and pathway analysis. This allows the integration of RPPA data with additional *omics* layers, e.g. transcriptomics data (messenger- and micro-RNAs), epigenetic modifications, or metabolomics data (Wrzodek *et al.*, 2013).

In a recent review, Waters *et al.* (2010) have investigated the integration of transcriptomics studies into the preclinical drug development process. The profiling of drug-induced changes in gene expression might allow an earlier assessment of potential undesired side effects in preclinical animal studies. The expression profiles that are obtained in these studies can be compared to reference datasets of chemicals with known side effects. With TG-GATEs (Uehara *et al.*, 2010) and DrugMatrix (Ganter *et al.*, 2005), two large datasets have been released to the public and are available for this purpose. Both datasets encompass gene expression profiles of rats after subchronic administration of carcinogenic and noncarcinogenic chemicals.

ToxDBScan uses a similarity scoring approach to identify well-characterized chemicals that induce gene expression changes similar to the observed gene expression profile (Römer *et al.*, 2014b). The effects of these most similar chemicals may provide leads for mechanistic analysis of the MOA and the potential side effects of the new compound. To this end, ToxDBScan calculates and reports a similarity score that is based on significantly deregulated genes. This similarity score can be used to rank the reference compounds by the similarity of the induced effects on gene regulation. ToxDBScan also performs a pathway enrichment analysis to aid the identification of possible MOAs. In our publication of ToxDBScan, we have shown that our similarity scoring approach is successful at identifying carcinogenic substances and have validated our results with external data for compounds that are not included in either TG-GATEs or DrugMatrix (Römer *et al.*, 2014b). A detailed description of ToxDBScan is given in Chapter 5.

4.2 Setup of the web platform

The ZBIT Bioinformatics Toolbox is hosted on a dedicated server, which is running a GNU/Linux server operating system. Users can access the ZBIT Bioinformatics Toolbox under the address `https://webservices.cs.uni-tuebingen.de` through their web browser. The server uses the Galaxy framework (Goecks *et al.*, 2010) to provide a user interface, serve static content, handle user requests (e.g., job submissions) and for data storage and database management. Galaxy is an open framework that was developed to provide web platforms for data intensive research, particularly next-generation sequencing projects in biomedical research. The Galaxy developers provide a public server that hosts the tools that are included with the framework. We set up our own Galaxy instance and use it to host the tools that were developed at our chair. The Galaxy framework is implemented mainly in Python, along with a medley of other components in JavaScript, HTML markup dialects, and others. Our own tools were implemented in either Java™ or the R programming language for statistical computing. We installed all requirements for Galaxy on the dedicated server. In addition, three nodes from our internal computation cluster have been assigned to the ZBIT Bioinformatics Toolbox to distribute the handling of web traffic and job execution. Each cluster node is running a GNU/Linux operating system. On the server, Apache2 and Python 2.7.3 were pulled from the main Ubuntu repository. We use the Oracle Grid Engine to handle the distribution and management of jobs on the small computation cluster. The Java™Runtime Environment (JRE, version 1.8.0) and R (version 3.2.3) (R Core Team, 2015) were installed on each cluster node to run the tools. All required third-party libraries were also installed on the cluster nodes. We integrated our tools into the Galaxy framework with a set of XML files that define the user interfaces and shell scripts that handle argument and output processing. Where necessary, we extended the tools to include a command line interface and respective argument handling. We enforce HTTPS and SSL encryption to secure all connections between users and the ZBIT Bioinformatics Toolbox. HTTPS

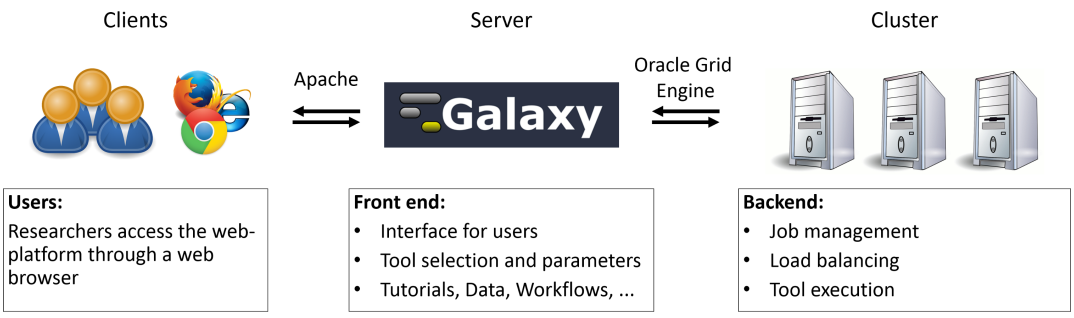


Figure 4.2: **General architecture of the ZBIT Bioinformatics Toolbox.** Clients access the platform through their browser. All in- and outgoing traffic is handled by Apache. On the server host, the Galaxy framework is used to provide the front end. Galaxy also handles user management, workflows, and persistent data storage. The Oracle Grid Engine distributes jobs to cluster nodes and manages the queue of running, waiting, and finished jobs. After the execution finishes, results are passed back along this command chain to Galaxy, which stores and displays results for clients. Figure from Römer *et al.* (2016b).

protects users and their data against unauthorized access and guarantees that users are communicating with the correct server. A schematic overview of the system is shown in Fig. 4.2.

The ZBIT Bioinformatics Toolbox provides several example workflows that represent use cases for each of the three branches of bioinformatics covered by the ZBIT Bioinformatics Toolbox. Workflows are analysis pipelines that can be built by using the output of one tool as the input of another tool. This is inspired by the UNIX philosophy: each tool does one task and does it well, which ensures modularity and reusability of tools and code, and by combining these simple tools, complex problems can be solved. Workflows can consist of two or more tools, and outputs of one tool can be used as input for several tools, e.g., for multiple visualization tools. The user can set the parameters that are used for each tool individually. Users can save workflows to reuse them for later, similar analysis or share them with other users who want to analyze similar problems. The preset parameters can be adjusted when using predefined workflows. This enables users to build a workflow that analyzes the data from their experiment and reuse the pipeline if they perform similar experiments or replications. An example of a common workflow is the combination of BioPAX2SBML and SBMLsqueezer in Path2Models project (Büchel *et al.*, 2013). An overview of the example workflows is given in Table A.1.

4.3 Use cases for the ZBIT Bioinformatics Toolbox

This section demonstrates the usage of the ZBIT Bioinformatics Toolbox with exemplary use cases for each of the three major categories: systems biology, transcription factor annotation, and expression data analysis. For these use cases, we used the predefined workflows to analyze real data obtained from public repositories. All data required for these use cases has been deposited in the ZBIT Bioinformatics Toolbox for reproduction and as an example for users.

4.3.1 Creation of full kinetic models from pathway maps

Curated pathway databases provide qualitative pathway maps that have been assembled manually from literature and primary research, e.g., KEGG (Kanehisa *et al.*, 2014), the Pathway Interaction Database (PID) (Schaefer *et al.*, 2009), or Reactome (Matthews *et al.*, 2009). The Path2Models project attempted to make these qualitative pathway maps usable for systems biology by converting these qualitative pathways to quantitative networks (Büchel *et al.*, 2013). For the Path2Models, Büchel *et al.* (2012) and Dräger *et al.* (2008) developed the tools BioPAX2SBML and SBMLsqueezer to allow the automated generation of SBML models that can be used for systems biology modeling from the BioPAX models that are available from the major pathway resources. In this use case, we demonstrate how to use the ZBIT Bioinformatics Toolbox to perform an extended version of the Path2Models approach. We use BioPAX2SBML, SBMLsqueezer, and SBML2LaTeX (which was not used in Path2Models) to convert a BioPAX model to the SBML format, add kinetic equations, and generate a human-readable report that can be used to inspect and verify the created SBML model.

For this example, we used the ceramide signaling pathway. Ceramides are sphingolipids from the family of waxy lipid molecules that are found in the cell membrane in high concentrations. Several researchers have demonstrated that ceramide signaling may be involved in differentiation, apoptosis, and programmed cell death (Haimovitz-Friedman *et al.*, 1997; Obeid *et al.*, 1993).

We downloaded the ceramide signaling pathway from PID in BioPAX and created a workflow that combines BioPAX2SBML, SBMLsqueezer, and SBML2LaTeX with default parameters. This workflow automatically pipes the output files of the previous tool to the next and generates a full kinetic model stored in the community standard SBML format and a human readable PDF report (see also Table A.1 and Fig. 4.3 (a)). The workflow performs three major steps: First, BioPAX2SBML converts the BioPAX file to SBML without loss of information (see Fig. 4.3 (b)). Second, SBMLsqueezer generates and adds kinetic equations for all reactions in the model (see Fig. 4.3 (c)). Third, SBML2LaTeX generates a human readable report as PDF. We used this workflow without changing the default options of the involved tools.

After the conversion of the qualitative model, the created SBML model contains 50 reactions with 93 involved molecules and 263 kinetic parameters. Any modeling soft-

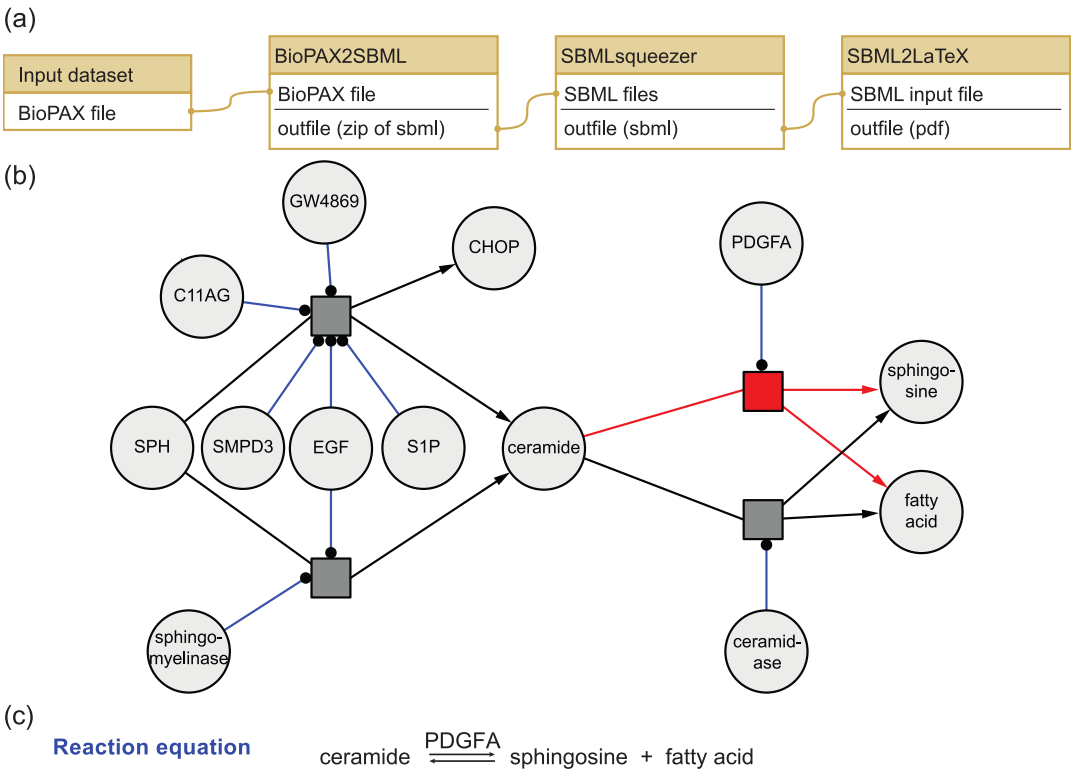


Figure 4.3: **Creation of a full kinetic model for the ceramide signaling pathway.** (a) Predefined Galaxy workflow for the creation of kinetic models from BioPAX files. BioPAX2SBML is used to convert the BioPAX encoded pathway to a draft SBML model. SBMLsqueezer infers reaction equations and kinetic rate laws for the relations defined in the resulting SBML model. SBML2LaTeX creates a human-readable report for model inspection to facilitate interpretation and curation. (b) This network represents a small part of the full ceramide signaling pathway that is involved in creation and degradation of ceramide. It contains four reversible reactions (dark gray squares) and 13 reactants. The black arrows indicate the participation of reactants in reactions. Blue lines indicate the enzymatic behavior of reactants. The reaction highlighted in red degrades ceramide to sphingosine and fatty acid and is catalyzed by Platelet-derived growth factor subunit A (PDGFA). We used CySBML (Konig *et al.*, 2012) to visualize the draft SBML model generated by SBMLsqueezer. (c) Reaction equation for ceramide degradation. SBML2-LaTeX creates reaction equations for all reactions in the PDF report. This reaction degrades ceramide to sphingosine and fatty acid and is catalyzed by the enzyme PDGFA. The reaction is also part of the subnetwork shown in (b) and is highlighted in red. Figure from Römer *et al.* (2016b).

ware that supports the SBML community standard and the SBML Level 3 `qual` package could then be used to perform further experiments with the ceramide signaling pathway. Both SBMLsqueezer and SBML2LaTeX provide options to further customize or improve the model and the report.

4.3.2 Identification of transcription factors and DNA binding domains

For this use case, we used NF-κB, which is a human TF that is present in almost all cell types and involved in the cell's response to stress induced by, for example, cytokines, radiation, or bacterial and viral antigens (Gilmore, 2006). NF-κB participates in the regulation of the immune response to infections and errors of its regulation have been linked to several adverse conditions such as cancer, autoimmune diseases, or immune system deficiencies.

We obtained the amino acid sequence of NF-κB from UniProt (ID: P19838) in the FASTA format and created a two-step workflow for this analysis: First, we used TFpredict to predict if NF-κB is a TF or not, along with its DBDs and structural superclass. Second, we used the results of the prediction with TFpredict as input for SABINE to infer the PFM of the DNA-binding site that is recognized by NF-κB. This workflow is available from the ZBIT Bioinformatics Toolbox as an example for users, along with the FASTA input file and additional sample sequences.

TFpredict correctly predicted NF-κB to be a TF (see Fig. 4.4a). TFpredict also predicted NF-κB to belong to the beta scaffold superclass and InterProScan identified four potential DBDs. Based on the identified DBDs and the structural superclass, SABINE was able to infer the putative PFM with medium confidence (see Fig. 4.4b and c). We compared the predicted PFM to the accepted PFM from the literature and found a good concordance. The PFM inferred by SABINE is `5'-GGRAANYCCC-3'`, the actual PFM is `5'-GGGRNYYYCC-3'`, where R represents a purine, Y a pyrimidine, and N any nucleotide (Wan and Lenardo, 2009).

4.3.3 Effects of drugs on protein expression

For this use case, we obtained a protein expression dataset that measured in the liver of rats which have been exposed subchronically to 11 NGCs, 2GCs, and 2 NCs, which is available from GEO under the accession number GSE53084 (Edgar *et al.*, 2002; Römer *et al.*, 2014a)). In this experiment, each group of three rats was administered with one substance or the corresponding vehicle for 14 days. Then, the rats were sacrificed, the liver was extracted, and protein expression was measured.

We used RPPApipe to assess the effects of NGCs on the protein expression in rat liver and compare it with the expression observed for GCs and NCs. To this end, we built an analysis workflow for a two-class experimental design, where for each substance the

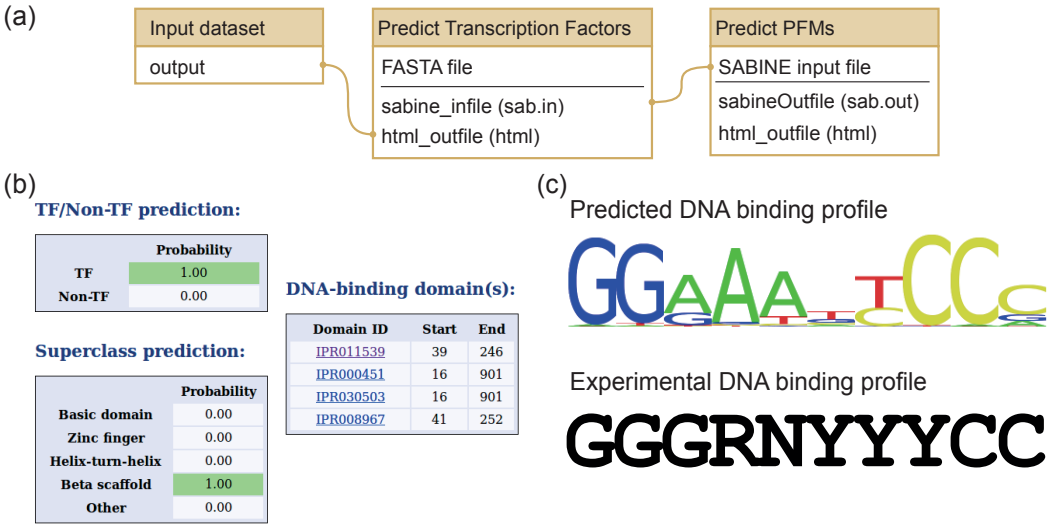


Figure 4.4: **Transcription factor prediction for human NF-κB with TFpredict and SABINE workflow.** (a) Predefined Galaxy workflow for transcription factor annotation. The input FASTA sequence file contains the protein sequence. TFpredict uses the sequence to predict if the protein is a transcription factor, infer its superclass, and detect DNA binding domains. SABINE uses the output of TFpredict to identify the DNA sequence that is bound by the transcription factor. (b) TFpredict output for NF-κB protein sequence. NF-κB was correctly classified as a transcription factor. TFpredict predicted the beta scaffold superclass and detected four DNA binding domains. (c) DNA binding profile predicted by SABINE. SABINE predicted a PFM with medium confidence. The predicted DNA binding profile shows good concordance with the consensus DNA binding profile established by Wan and Lenardo (Wan and Lenardo, 2009). In the consensus sequence, R represents a purine, Y a pyrimidine, and N any nucleotide.

group of rats receiving the vehicle is used as the control for the group of rats that received the actual substance. The workflow has three phases with a total of 12 tools and requires as input the observed protein expression in the animals (also referred to as samples) and a class definition file, which defines the relations of treatments and controls (see Fig. 4.5a). First, during the preprocessing, each sample is assigned to a treatment group according to the defined relations. Here, we omitted scaling and log-transformation, as the data was already quantile normalized. In the preprocessing, RPPApipe also fetches additional information from public databases, e.g., protein descriptions or alternative identifiers. Second, RPPApipe calculates fold changes and *p*-values to determine which proteins are differentially expressed between the treated samples and the controls. We used limma (Smyth, 2005) to identify differentially regulated proteins and corrected *p*-values for multiple testing with the Benjamini-Hochberg correction (Benjamini and Hochberg, 1995). Third, RPPApipe generates several plots to visualize the results of the experiment, e.g., volcano plots for differential regulation and modification, and KEGG (Kanehisa and Goto, 2000) pathway profiles (see Fig. 4.5b). These visualizations allow an easy interpretation of the general effects that we observed in the experiment.

The clustering analysis shows that there is a distinct difference in the protein expression in rat liver after administration with the two NCs, Nifedipine and Cefuroxime, and the carcinogenic substances, as can be seen in Fig. 4.5c. We also see a lesser distinction between the NGCs and the two GCs, C.I Direct Black 38 and Dimethylnitrosamine. This suggests that carcinogenic substances can be identified in short-term assays, e.g., by measuring protein expression after subchronic exposure to the drug candidate (Römer *et al.*, 2014a).

4.4 Related web platforms

In the field of systems biology, the ZBIT Bioinformatics Toolbox includes the tools BioPAX2SBML, SBMLsqueezer, SBML2LaTeX, and ModelPolisher. The SBML community website[1] provides the SBML Software Guide, which is actively maintained and lists software and tools that use or support the SBML format. The SBML Software Guide includes web platforms in several categories such as SBML editing (semanticSBML, (Krause *et al.*, 2010)), visualization (PATIKAweb, (Dogrusoz *et al.*, 2006)), and annotation (MetaNetX, (Ganter *et al.*, 2013)). However, of the tools listed, none provides the functionality of the tools that we provide through the ZBIT Bioinformatics Toolbox. In the category of BioPAX converters, BioPAX2SBML has only one alternative: SyBiL, which is not available as a web platform and requires installation, along with mandatory third-party libraries. While there are other SBML converters that are also available from web platforms, none is able to convert BioPAX models to SBML. For example, the System Biology Format Converter Online[2] supports BioPAX conversion only to GPML,

[1] `http://sbml.org`, accessed Feb. 22, 2016

[2] `https://www.ebi.ac.uk/biomodels/tools/converters/`, accessed Feb. 22, 2016

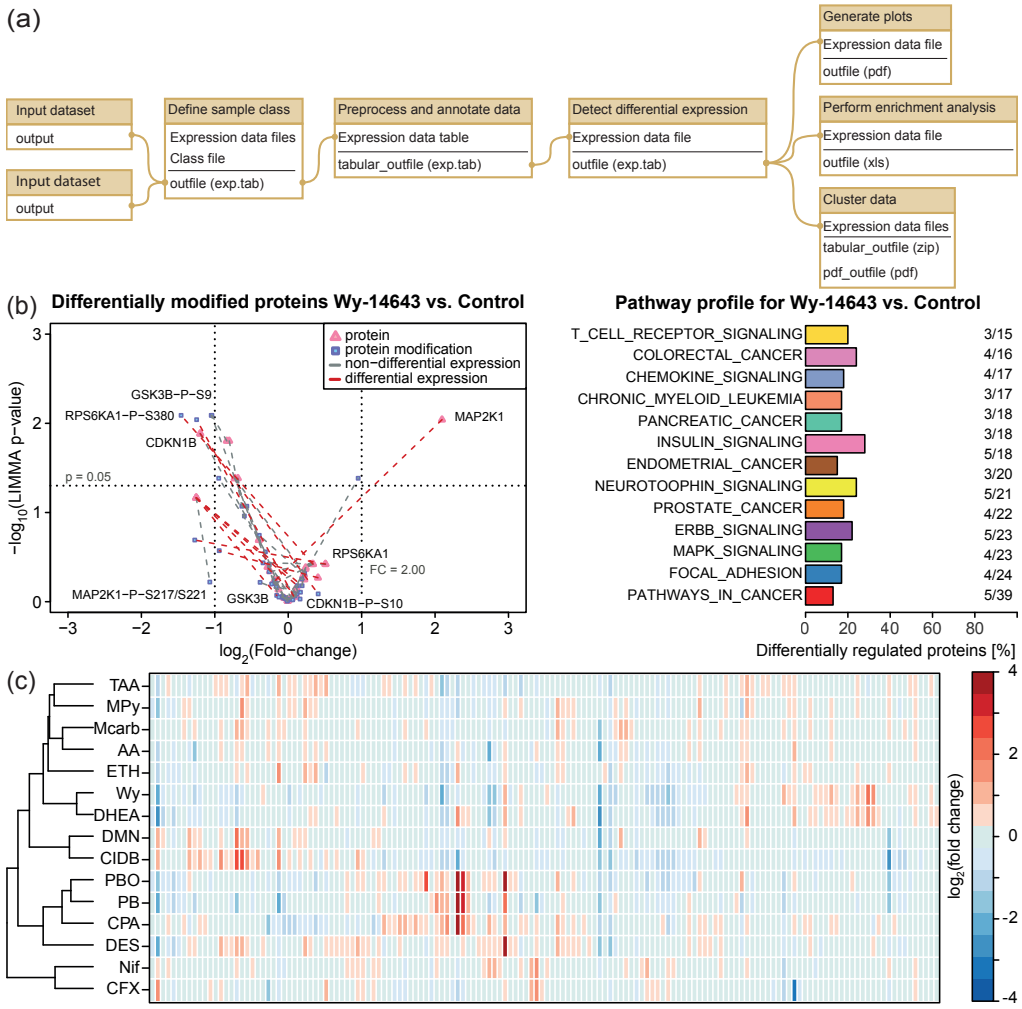


Figure 4.5: **Analysis of the effects of drugs on protein expression with RPPApipe.** (a) Predefined Galaxy workflow for RPPA data analysis. Two input files are required: a CSV file containing the RPPA expression values and a class file, which defines the relations between samples. First, the data is normalized and annotated. Second, differential expression of proteins is determined. Third, various plots are generated, and pathway enrichment and clustering are performed. (b) Volcano plot and pathway profile for RPPA data. These example plots were generated using a dataset for effects of drug exposure on the protein expression in rat liver. Several proteins are differentially modified after treatment with Wy-14643, a NGC (left). Differentially regulated proteins have been mapped to KEGG pathways to identify potential deregulation on pathway level (right). (c) Clustering of drugs by effects in rat liver. All 15 drugs in the dataset were clustered by their protein expression profiles. The two NCs (Nif, CFX) formed a separate cluster from the carcinogens. The two GCs (CIDB, DMN) formed a cluster within the carcinogens. The NGCs formed several clusters. Figure from Römer *et al.* (2016b).

not to SBML. Currently, none of the other three systems biology tools has a direct competitor listed in the SBML Software Guide, neither as stand-alone software nor as a web platform. With SBML2TikZ (Shen *et al.*, 2010) there is a tool that provides additional functionality that might be used together with SBML2LaTeX, e.g., to render the SBML model in addition to generating a human-readable summary. However, SBML2TikZ is no longer available as a web tool and must be installed locally.

For transcription factor analysis, the ZBIT Bioinformatics Toolbox provides TFpredict and SABINE. The experimental methods that researchers have traditionally used to identify were time-consuming and expensive. This has stimulated the development of a number of *in silico* methods in the last decade, which use computational models or machine learning. Among the published methods for identification of DNA-binding proteins are also web platforms. The two most notable examples are iDNA-Prot|dis by Liu *et al.* (2014) and nDNA-Prot by Song *et al.* (2014), which are specifically designed for the prediction of DNA binding based on the amino acid sequence. Both iDNA-Prot|dis and nDNA-Prot are accessible through web interfaces that consist of a paste box for the DNA sequence without further options or much documentation. Both tools only provide the prediction, DNA-binding or not, without further information. In contrast, TFpredict predicts if the protein is a TF or not and also infers the superclass and the DBDs with InterProScan. Both iDNA-Prot|dis and nDNA-Prot have been reported to outperform their competitors, but have not been compared to each other, nor to TFpredict. An additional advantage of TFpredict is the compatibility with SABINE, which allows further analysis, whereas the other two web tools only provide raw prediction results. For the PFM prediction with SABINE, we have not found any competing web platforms.

For analysis of expression data, the ZBIT Bioinformatics Toolbox provides RPPApipe for protein expression data and ToxDBScan for gene expression data. Wachter *et al.* (2015) have recently reviewed the tools which are available for the analysis of RPPA data. In their list, they mention only two web platforms: RPPApipe and Miracle by List *et al.* (2014). All other tools discussed by Wachter *et al.* (2015) are either Excel macros or R packages. For this reason, we compared RPPApipe only to Miracle, which in theory also provides a web platform. However, at the present, only the source code is available on GitHub and no official Miracle web server exists. Users are required to set up a custom server, which requires infrastructure for hosting a web server and advanced programming and administration knowledge. Thus, RPPApipe is currently the only analysis pipeline for RPPA data that is freely available as a web platform. In particular, RPPApipe provides specific analysis methods and visualizations for RPPA data analysis, e.g., the analysis of differential modification of proteins. For the more general case of finding patterns in the RPPA expression data without addressing the RPPA specific questions, users could use web platforms designed for gene expression analysis. For example, PaGeFinder (Pan *et al.*, 2012) provides pattern analysis for user submitted expression data and PaGenBase provides a database of pattern genes in a number of model organisms (Pan *et al.*, 2013). These web platforms can be used to identify genes that are specifically expressed in certain conditions, e.g., to identify spatiotemporal patterns in sequential gene expression

experiments.

ToxDBScan is a similarity search engine for gene expression patterns which scans the two largest databases for the effects of NGCs on gene expression: TG-GATEs and DrugMatrix. Currently, two other tools provide related functionality: Toxygates by Nystrom-Persson *et al.* (2013) and LTMap by Xing *et al.* (2014). Toxygates is a data portal which provides exploration tools for the TG-GATEs data, allows compound ranking by gene expression and links expression data with pathology reports (Nystrom-Persson *et al.*, 2013). However, Toxygates does not provide similarity search based on differentially regulated genes provided by the user. LTMap performs similarity ranking based on user-submitted probe lists in TG-GATEs data, but does not offer any additional analyses (Xing *et al.*, 2014). In contrast, ToxDBScan performs pathway enrichment analysis for the submitted data and renders heat maps that enable researchers to visually inspect the similarity of the gene expression profiles. Also, at the time of writing, the LTMap server was not available and we received no response from the developers when we informed them of this issue. A definitive advantage of ToxDBScan in comparison with Toxygates and LTMap is the integration of the DrugMatrix data, which almost doubles the number of compounds available for similarity search.

4.5 Summary and conclusions

In this chapter, we presented the ZBIT Bioinformatics Toolbox, which is a web platform for bioinformatics tools. In total, the ZBIT Bioinformatics Toolbox encompasses eight tools that have been developed at the chair of Cognitive Systems at the University of Tuebingen. These tools solve problems in three branches of bioinformatics: systems biology, expression data analysis, and transcription factor analysis. In systems biology, the focus of the ZBIT Bioinformatics Toolbox is on SBML conversion and model development. BioPAX2SBML is a converter that translates biological networks from BioPAX to SBML. SBMLsqueezer infers kinetic rate laws for dynamic simulation of SBML models. SBML2LaTeX generates human-readable reports from the XML-based SBML models for error-checking and model presentation. ModelPolisher augments SBML models with additional annotations and information from the BiGG models knowledge base. For expression data analysis, the ZBIT Bioinformatics Toolbox focuses on protein array data processing and gene expression database screening. RPPApipe offers a pipeline for RPPA data processing, annotation, analysis, and visualization. ToxDBScan performs large scale similarity screening for gene expression profiles in two large toxicological databases. For transcription factor annotation, TFpredict predicts if a protein is a TF or non-TF based on its amino acid sequence. SABINE uses the results of TFpredict to infer the DNA motif that is recognized by the identified TFs. The tools in the ZBIT Bioinformatics Toolbox have been used by researchers to gain new knowledge in systems biology (Pathak *et al.*, 2013; Gupta and Misra, 2013; Schröder *et al.*, 2011) and are adopted in established databases (Li *et al.*, 2010).

The ZBIT Bioinformatics Toolbox is a free resource for life science researchers who want to use our bioinformatics tools and will be maintained and extended if more tools are developed. The most recent addition to the ZBIT Bioinformatics Toolbox was ModelPolisher, which was added in early 2016. Because it is a web platform, it can be accessed through the browser from any device without any hardware restrictions. The only requirements for using the ZBIT Bioinformatics Toolbox is a browser and Internet access. Therefore, in theory, biologists could use mobile devices to submit jobs while in the lab. Another advantage of the ZBIT Bioinformatics Toolbox is that it requires no technical knowledge for use, in contrast to the original tools, which would require the installation of third-party software or knowledge of command line usage. We used the Galaxy framework to implement our web platform. The Galaxy framework provides an established interface that is familiar to many biologists who already used other bioinformatics tools on other web platforms. The framework was developed for scientific software and provides excellent support for storage of results, scalability, and reproducibility in concordance with the requirements for scientific software. In addition, it allows the creation of analysis pipelines, so-called workflows, which combine multiple tools to perform complex analysis without much effort. The ZBIT Bioinformatics Toolbox provides examples and tutorials for all tools along with links to documentation. We have compared our web platform with other web tools that were created to solve similar tasks and showed that the ZBIT Bioinformatics Toolbox either provides added value for researchers or provides unique tools that are not available from other web platforms. We hope that the ZBIT Bioinformatics Toolbox will facilitate the usage of our bioinformatics tools for life science researchers and thus further their research and enable new insights.

Chapter 5

Similarity screening for characterization of drug candidates

Chapter 1 introduced the MARCAR project, which is concerned with the identification of early biomarkers for nongenotoxic carcinogenesis. While genotoxic drugs are eliminated at the beginning of the development process by *in vitro* assays like the Ames test (Ames *et al.*, 1975), nongenotoxic carcinogens (NGCs) cannot yet be reliably detected by *in vitro* assays (Jacobs, 2005). Drugs that are to be approved for chronic human administration have to provide an estimation of carcinogenic risk. The carcinogenic risk is commonly assessed using the LRB, a 2-year, *in vivo*, chronic administration assay with rodents, typically rats or mice. If carcinogenic effects are observed after the completion of these long-term rodent bioassays, the market introduction of new, potentially important medicines for patients may be delayed. Often, these long-term bioassays are performed as one of the last steps in the safety assessment, after the first clinical phases are already running or even completed, require 5 years until completion, and cost up to 2 million US dollars (Johnson, 2012). However, due to a high rate of false positives and background tumor incidences in control animals, these long-term rodent bioassays have drawn a lot of criticism (e.g., Johnson, 2012; Cohen, 1995; Jacobs, 2005). Other problems include appropriate selection of the administered dose and that the observed endpoints do not provide mechanistic insight into the process of cancer formation.

To address these problems and provide more informative short-term assays, many researchers have proposed toxicogenomics, which employs *in silico*, *in vitro*, and *in vivo* methods. Currently, the most promising approach in toxicogenomics are short-term *in vitro* or *in vivo* rodent assays that use microarrays or RNA-seq to profile global gene expression, which is then analyzed with supervised or unsupervised machine learning and pattern recognition methods (see Chapter 3 and the review by Waters *et al.* (2010)). The studies which investigated these short-term rodent assays are often based on a small number of drugs and use only the data collected in a single study and by a single group. Frequently, competing groups also use costumized profiling technologies (e.g., different microarray platforms) and only report chip-specific identifiers, which further complicates the interpretation of the study and the comparison with external data. Due to the heterogeneous modes of action of nongenotoxic substances, the small sample sizes may

be particularly problematic. For this reason, two large databases have been established through the cooperation of national institutes and companies: TG-GATEs (Uehara *et al.*, 2010) and DrugMatrix (Ganter *et al.*, 2005), which also used the same profiling technology. The large number of compounds that was used in comparison with previous studies may allow an improved analysis of common MOAs and similarities of nongenotoxic substances. Using the gene expression profiles in the two databases, toxicologists could predict the carcinogenicity of new compounds and identify similar compounds to extrapolate information, e.g., potential side-effects. At the time of writing, to our knowledge, only one study had assessed cross-prediction of TG-GATEs and DrugMatrix (Gusenleitner *et al.*, 2014) and no tools were available that made use of both databases.

This chapter presents ToxDBScan, a web application that searches TG-GATEs and DrugMatrix for substances that induce gene expression patterns which are similar to the patterns induced by new substances. We designed ToxDBScan to be accessible to all researchers, independent of technical background or the profiling technique that they used to profile gene expression patterns. ToxDBScan uses a newly developed similarity scoring system based on a modified Tanimoto similarity index and provides the user with an HTML report that can be viewed in most modern web browsers. The ToxDBScan web application is part of the ZBIT Bioinformatics Toolbox, which was described in Chapter 4. ToxDBScan only requires the up- and downregulated genes to identify similar compounds and thus does not oblige the user to provide any potentially confidential data, e.g., the chemical structure, the experimental setting, or microarray data. This allows a quick and easy identification of potentially similar compounds for further mechanistic analysis, assessment of their hepatocarcinogenic potential, or MOA discovery. The content of this chapter was published in the *International Journal of Molecular Sciences* under the title "ToxDBScan: Large-scale similarity screening of toxicological databases for drug candidates" (Römer *et al.*, 2014b).

5.1 Data resources

We used data from three public databases in the development of ToxDBScan: the Carcinogenic Potency Database (CPDB), TG-GATEs, and DrugMatrix. The CPDB provides information on the outcome of long-term rodent carcinogenicity assays for more than 1000 compounds. TG-GATEs and DrugMatrix provide profiles of the gene expression in the liver of male rats after subchronic administration of hundreds of compounds. The combination of results from long-term assays as a gold-standard carcinogenicity annotation and gene expression changes after subchronic administration provides users with a lot of additional information that can help in the characterization of compounds for which they have obtained gene expression profiles. This section describes these three major data resources for ToxDBScan and explains their relevance in toxicogenomics in general.

5.1.1 The Carcinogenic Potency Database

The CPDB was established by Gold *et al.* (1984) as a database for standardized results of animal bioassays. Currently, the CPDB is hosted by the U.S. National Library of Medicine as part of the TOXNET[1], a resource for searching toxicology databases which list hazardous substances, literature on their biochemical effects, and more (Fitzpatrick, 2008). Since 2011, the CPDD is no longer updated following the death of its director Lois Swirsky Gold, who initially established the CPDB in 1984. Nevertheless, the CPDB is an invaluable resource for toxicogenomics because it is currently the only public machine-readable database for carcinogenicity and genotoxicity annotations.

The CPDB records results of chronic, long-term *in vivo* animal cancer tests that were reported in the general literature and by the National Cancer Institute until 2004. As of its last update in 2005, the CPDB encompasses both positive and negative results of 6540 bioassay experiments for 1547 chemicals performed in rats, mice, hamsters, dogs, and nonhuman primates. It standardizes the variations in protocols, nomenclature, and information provided by the authors of the original literature and provides an easy and machine-readable format. Most importantly for the function of ToxDBScan, the CPDB lists tumor occurrence by organism, sex, and target organ. This allows ToxDBScan to provide site-specific information, e.g., only results which are found in the liver of male rats. In addition, the CPDB provides the outcome of an auxotroph-based Ames test for many chemicals.

We used the CPDB to assign hepatocarcinogenicity for compounds as follows: First, a compound was classified as a carcinogen (C) if there was at least one positive experiment which listed the liver as a target organ. Otherwise, the compound was classified as an NC. Second, we assigned genotoxicity through the recorded result of the Ames test. If a positive Ames test is recorded, the compound was classified as genotoxic, and if a negative Ames test is recorded, it was classified as a nongenotoxic. Finally, we assigned one of three classes, based on the genotoxicity and hepatocarcinogenicity: GC if the compound is genotoxic and hepatocarcinogenic, NGC if the compound is not genotoxic, but hepatocarcinogenic, or NC if the compound is not hepatocarcinogenic, regardless of genotoxicity. Compounds are considered unclassified if no carcinogenicity tests is recorded in the CPDB or if they are hepatocarcinogenic but have no recorded Ames test results. To annotate unclassified compounds, we used the annotations by Uehara *et al.* (2011), who provide classifications for the compounds included in TG-GATEs.

5.1.2 Open TG-GATEs

The TG-GATEs database was established by a consortium of the Japanese government and several Japanese pharmaceutical companies (Takashima *et al.*, 2006). In 2010, it was released to the public under the name Toxicogenomics Project-Genome Assisted Toxicity

[1] `http://toxnet.nlm.nih.gov/cpdb/`, accessed March 8, 2016

Evaluation System (TG-GATEs, Uehara *et al.*, 2010). TG-GATEs was the first public toxicogenomics project that established an extensive database of gene expression profiles in the liver of male rats after subchronic administration of drugs and other compounds.

In total, TG-GATEs contains microarray data for 160 chemicals, which include carcinogenic substances as well as approved drugs. These 160 chemicals were administered to male Sprague-Dawley rats for the subchronic *in vivo* assays and cultured human and rat hepatocytes for *in vitro* assays. In the *in vivo* experiments, liver and kidney were removed and profiled for gene expression after a defined exposure duration. Affymetrix Rat Genome 230 2.0 microarrays were used to profile the gene expression for *in vivo* experiments and *in vitro* rat hepatocyte experiments. Each compound was administered in three dose levels: low, middle (3-fold low dose), and high dose (10-fold low dose). Each dose level for each compound was further profiled at different time points: either 3, 6, 9, and 24 hours after a single administration, or 4, 8, 15, or 29 days after repeated daily administration of the compound. Each condition, i.e., each combination of chemical, dose level, and duration, was performed in triplicates. In total, TG-GATEs contains data for 14,143 microarrays, which provide data for 3,528 different combinations of chemical, dose level, and exposure duration and the corresponding, time-matched controls. The raw data for these microarrays was deposited at ArrayExpress (Kolesnikov *et al.*, 2015) under the accession number E-MTAB-800.

For ToxDBScan, we used only data that was obtained by *in vivo* experiments through the profiling of gene expression in the liver of male rats. Through CPDB and Uehara *et al.* (2011), we were able to annotate 123 compounds, which account for 2,768 conditions. We obtained the raw microarray data from the ArrayExpress FTP server and preprocessed the data using RMA normalization implemented in the `affy` package for R/Bioconductor (Gentleman *et al.*, 2004; Gautier *et al.*, 2004).

5.1.3 DrugMatrix

The DrugMatrix data is available from the Gene Expression Omnibus (GEO, Barrett *et al.* (2004)) under the accession number GSE57822. In total, DrugMatrix contains expression profiles for 612 compounds, of which in 2005, 460 were approved drugs, 25 were withdrawn drugs, and 127 were reference compounds or compounds of toxicological interest (Ganter *et al.*, 2005). The expression profiles were compiled from seven different tissues, e.g., liver, kidney, and heart, taken from male Sprague-Dawley rats after subchronic administration. Codelink microarrays were used for expression profiling in the initial studies. Later, the tissue samples were profiled with the Affymetrix Rat Genome 230 2.0 Array, which was also used by TG-GATEs. DrugMatrix contains profiles after both single (six hours and one day) and repeated (three and five days) administrations, which leads to a total of 3,200 different drug-dose-time-tissue combinations, for which time-matched controls are available.

Again, for ToxDBScan we used only the expression profiles that were obtained from the liver of the male rats. Thus, 1,939 expression profiles are available for 200 of the 612

compounds, which accounts for 654 of the 3,200 drug-dose-time-tissue combinations. These were performed mostly in triplicates with time-matched controls, however, for some conditions only one or two replicates were available. Through CPDB, we were able to annotate 132 of the 200 compounds, which account for 440 of the 654 conditions. If compounds with missing annotation were also included in the TG-GATEs database, annotations from Uehara *et al.* (2011) were added where possible. We obtained the RMA normalized data from the DrugMatrix FTP server[2].

5.1.4 Comparison of Open TG-GATEs and DrugMatrix

TG-GATEs and DrugMatrix are the two biggest expression profile databases available for toxicological research and are also among the largest resources in microarray research in general. At the time of writing, ArrayExpress listed only one study, a comprehensive human expression map collected from multiple resources, which was larger in terms of the number of microarrays used (16,241 for TG-GATEs and 10,899 for DrugMatrix[3]). Both databases are publicly available and contain more than 100 compounds, far more than most other toxicogenomics studies used. More importantly, both databases were profiled using the same microarray platform (Affymetrix Rat Genome 230 2.0 Array), which greatly simplifies the comparison of expression profiles, because no identifier mapping is required. We used the CAS numbers to identify compounds that are contained in both TG-GATEs and DrugMatrix and found an overlap of 51 compounds.

However, there are differences in the chosen exposure durations and dose levels between the two databases. For TG-GATEs, the highest dose was selected as the maximum dosage that is considered acceptable for one month of repeated dosing (Uehara *et al.*, 2010). In DrugMatrix, the dose levels are based on estimates of the maximum tolerated dose or the fully effective dose, which were based on previous dose finding studies and literature research (Giffin *et al.*, 2008). As a consequence, the administered dosage is higher in DrugMatrix for almost all of the 51 compounds that are present in both databases. Also, the longest exposure duration in DrugMatrix, 5 days, is much shorter than the 29 days used in TG-GATEs, which is most likely also a result of the different strategies of dosage selection. According to Gusenleitner *et al.* (2014), the doses used in TG-GATEs are considered suitable for the LRB and thus could be more useful for the prediction of the bioassay result, whereas due to the high dosages used in DrugMatrix, some chemicals which are not toxic in conventionally applied doses might show strong toxic effects.

[2] `ftp://anonftp.niehs.nih.gov/drugmatrix/`, accessed March 10, 2016
[3] `http://www.ebi.ac.uk/arrayexpress/browse.html`, accessed March 10, 2016

5.1.5 Validation dataset

In addition to TG-GATEs and DrugMatrix, we obtained a third dataset of gene expression profiles after subchronic administration of pharmaceutical substances. This dataset is not included in either TG-GATEs or DrugMatrix but was established by the Bayer AG and published as part of the MARCAR project (Römer *et al.*, 2014a). We use this dataset as an external validation set for our new similarity coefficient and the prediction method that we describe below. It is available from GEO under the accession number GSE53082. The validation dataset contains 15 compounds, of which 11 are NGCs, 2 are GCs, and 2 are NCs.

Whereas the expression data is not included in either TG-GATEs or DrugMatrix, independent profiles for 10 of the substances in the validation dataset are encompassed by one or both databases. The expression profiles in the validation dataset were generated using a different microarray platform, the Affymetrix Rat Genome 230a Array, which contains only a subset of the probes on the Affymetrix Rat Genome 230 2.0 Array used for TG-GATEs and DrugMatrix. An overview of the 15 substances is provided in Table A.2. We obtained the RMA normalized data from GEO and applied all additional processing steps as described for TG-GATEs and DrugMatrix.

5.2 Similarity scoring for gene expression profiles

Finding similarities between gene expression profiles is a common problem for microarray experiments. Several similarity measures have been proposed, Pearson or Spearman rank correlation and the Euclidean distance are the most common, but others have also been investigated (Yona *et al.*, 2006). These measures share a common problem: they require numerical values, such as fold changes or ranks, for the same entities (e.g., probe sets or gene symbols). In microarray experiments, a systematic bias between different laboratories may lead to varying dynamic ranges that could confound these similarity measures.

The large number of different microarray platforms that are available also requires the mapping of identifiers or restriction to a common set of identifiers. Furthermore, the number of genes that are deregulated in a certain experiment is usually expected to be small compared to the number of genes on the microarray chip. Thus, small, random variations may prevent the detection of similar profiles if, for example, the Euclidean distance is used. Because the number of differentially expressed genes (DEGs) is generally expected to be small, we propose to use sparse gene fingerprints and similarity coefficients for sets to measure the similarity of these fingerprints.

5.2.1 Gene expression fingerprints

So-called fingerprints are often used as feature representation of chemicals in chemoinformatics. A well-known example is the extended connectivity fingerprint, which represents a chemical compound as a set of its contained substructures (Rogers and Hahn, 2010). Here, we define a gene expression fingerprint as the set of genes (identified by their official gene symbol) that is found differentially expressed in a specific condition. In addition, we retain information on the gene's direction of regulation, such that we know whether the gene was up- or downregulated for each element of the set. The thresholds used to select DEGs are dependent on specific experiments and are based mainly on two criteria: intensity ratios and p-values. Intensity ratios are calculated by dividing the observed average intensity in treated conditions by the observed average intensity in control conditions. These intensity ratios are also called fold changes. Here, whenever we refer to fold changes, we mean the base 2 logarithmized intensity ratio. The p-values are obtained by statistical analysis of the intensity ratio and the observed variations in the intensities. Several methods have been proposed to calculate p-values for microarrays, e.g., simple t-tests or more sophisticated approaches using moderated t-statistics (Smyth, 2005). However, most methods require at least three replicates of both treated and control conditions to calculate p-values. In general, both approaches are combined such that genes are considered to be differentially expressed if the observed intensity ratio is large and the p-value is below a chosen false-discovery rate, e.g., less than 5%.

Here, we used only the fold changes to identify DEGs because for many conditions in DrugMatrix only one or two replicates were available. We also mapped probe sets to gene symbols using the `biomaRt` package (Durinck *et al.*, 2009) and summarized all probe sets that are assigned to the same gene symbol. Finally, we used two fold change thresholds to select DEGs in TG-GATEs and DrugMatrix and extract the gene expression fingerprints for each condition.

5.2.2 Tanimoto similarity coefficient and Jaccard index

Gene expression fingerprints are sets of DEGs identified from gene expression profiles. For this reason, we need a similarity measure for sets to calculate the similarity between gene expression fingerprints. The Tanimoto coefficient is a well-established similarity measure for sets and is used, for example, in chemoinformatics for substructure and similarity searching. The Tanimoto coefficient was derived from the Jaccard index (Levandowsky and Winter, 1971), which is also called Jaccard similarity coefficient and uses the Jaccard distance, a metric for calculating the distance between two arbitrary sets. It is defined as the ratio of the number of elements in the overlap of two sets A and B and the number of elements in the union of the two sets and can be calculated as

$$J = \frac{|A \cap B|}{|A \cup B|} \tag{5.1}$$

This is equivalent to calculating

$$J = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{5.2}$$

which does not require the union of the sets A and B. The Tanimoto coefficient T is the equivalent of the Jaccard index defined on binary vectors $X, Y \in \{0,1\}^n$:

$$T = \frac{\sum_{i=1}^{n} X_i \wedge Y_i}{\sum_{i=1}^{n} X_i \vee Y_i} \tag{5.3}$$

By reformulating the sets A and B, the Tanimoto coefficient can be used to score arbitrary sets. First, define a vector V that enumerates all entries of the sets A and B. Then, let X and Y be vectors of the same length as V, where X_i and Y_i are 1 if V_i is in the set A or B, respectively, or 0 otherwise. The Tanimoto coefficient of X and Y will then be equal to the Jaccard index of A and B.

5.2.3 Novel similarity coefficient for gene expression fingerprints

The gene expression fingerprints that we defined above contain information on the direction of regulation of each gene. For this reason, our gene expression fingerprints are not binary, but ternary, i.e., we have three states for each gene: "1" means the gene is upregulated, "0" means the gene is not differentially regulated, and "-1" means the gene is downregulated. We modified the Tanimoto coefficient to calculate the similarity of ternary vectors as follows: For two gene fingerprints $X, Y \in \{-1,0,1\}^n$, where n is the number of measured genes, the similarity score S is defined as:

$$S = \frac{\sum_{g=1}^{n} \delta(X_g, Y_g)}{\sum_{g=1}^{n} |X_g| + |Y_g| - \delta(|X_g|, |Y_g|)} \tag{5.4}$$

where $\delta(x,y)$ is defined as:

$$\delta(x,y) = \begin{cases} 1, & x = y \neq 0 \\ 0, & \text{else} \end{cases} \tag{5.5}$$

With our modified score S, we can calculate the similarity of gene expression profiles by extracting their gene expression fingerprints and calculating the similarity of the fingerprints. For binary vectors our modified Tanimoto coefficient S and the original Tanimoto coefficient T are equal.

In previous experiments, we observed that the administration of pharmaceutical compounds often induces an unspecific response, such that a number of genes are found to be differentially regulated for many compounds, regardless of their toxicological properties. To account for this unspecific response, we introduced a weight factor w for each gene, which is dependent on the observed frequency of deregulation. A higher frequency of deregulation in the database implies that the gene is part of an unspecific response to drug administration and may not be relevant for the assessment of potential drug effects and thus means that the gene should get a lower weight. In contrast, genes that are only found to be differentially regulated in few conditions are given a higher weight, as their deregulation is considered to be more specific for certain drug effects. We calculate the weight for each gene as

$$w_g = -\log_{10} \frac{\sum_{c \in C} |c_g|}{N} \tag{5.6}$$

where N is the number of compounds in the database and C is the set of gene fingerprints of the database compounds, i.e., c_g is 1 if gene g is upregulated in compound c, -1 if g is downregulated, and 0 if g is not deregulated. Thus, w_g corresponds to the negative decadic logarithm of the probability of observing deregulation of gene g when randomly choosing a condition from the database. This approach is inspired by information theory, where this concept of information content or self-information (Cover and Thomas, 2006) is used to distinguish between relevant signals and noise. It should be noted that the calculation of the information content of genes depends on the database and requires the inclusion of reference compounds and compounds which are not toxic. By including the weight, the modified coefficient is computed as

$$S = \frac{\sum_{g=1}^{n} w_g \delta(X_g, Y_g)}{\sum_{g=1}^{n} w_g \left(|X_g| + |Y_g| - \delta(|X_g|, |Y_g|)\right)} \tag{5.7}$$

5.3 Evaluation of the new similarity coefficient

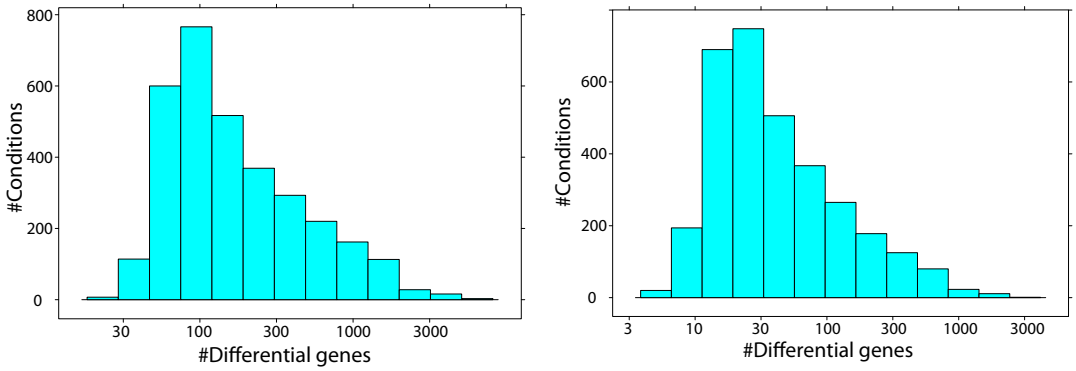
The following section describes the validation of the gene expression fingerprints and the proposed similarity coefficient. For the validation, we have used a dataset that is not contained in either TG-GATEs or DrugMatrix and was generated by an independent group in a different laboratory using a different microarray platform.

5.3.1 Gene expression fingerprint extraction

The first step of the evaluation was the extraction of gene expression fingerprints for all conditions (compound-dose-time combinations) in TG-GATEs and DrugMatrix. We extracted the fingerprints as described above by calculating intensity ratios (treated to

control animals) using the RMA normalized data and subsequently filtering for genes that are at least 1.5-fold (low threshold) or 2-fold (high threshold) up- or downregulated. We selected these values because they are established as common intensity ratio thresholds in gene expression analysis. We did not include a p-value threshold because many conditions (and in some cases controls) in DrugMatrix were not performed in triplicates. Using this procedure, we found at least one up- or downregulated gene in each condition in both TG-GATEs and DrugMatrix. The distribution of fingerprint sizes (i.e., the number of up- or downregulated genes for a condition) is shown in the histograms in Fig. 5.1 for the low and high threshold.

The lower intensity ratio threshold lead to gene fingerprint sizes between 23 and 6,525 genes, with a median fingerprint size of 131 genes. On average, we observed smaller fingerprints in TG-GATEs with a median size of 111 genes and larger fingerprints in DrugMatrix with a median size of 603 genes. For the higher intensity ratio threshold, the fingerprints are smaller, as would be expected. The fingerprint size ranges from 5 to 3,224 genes, and the median fingerprint size was 32 genes. Again, fingerprints in TG-GATEs are on average smaller than in DrugMatrix, with a median fingerprint size of 27 genes in TG-GATEs compared to 152 in DrugMatrix. The higher compound doses that were used in DrugMatrix are the most likely reason for the higher numbers of DEGs in DrugMatrix samples. However, the shorter exposure time or inter-laboratory differences could also lead to different dynamic ranges for the experiments.



(a) 1.5-fold up- or downregulation (low)

(b) 2-fold up- or downregulation (high)

Figure 5.1: **Gene expression fingerprint sizes for different intensity ratio thresholds.** Gene expression fingerprints were extracted by filtering genes by the ratio between the observed normalized intensities in treated and control samples. These histograms show the distribution of the size of the fingerprints for all conditions in TG-GATEs and DrugMatrix. Median fingerprint sizes are 131 genes for the lower threshold (a) and 32 genes for the higher threshold (b).

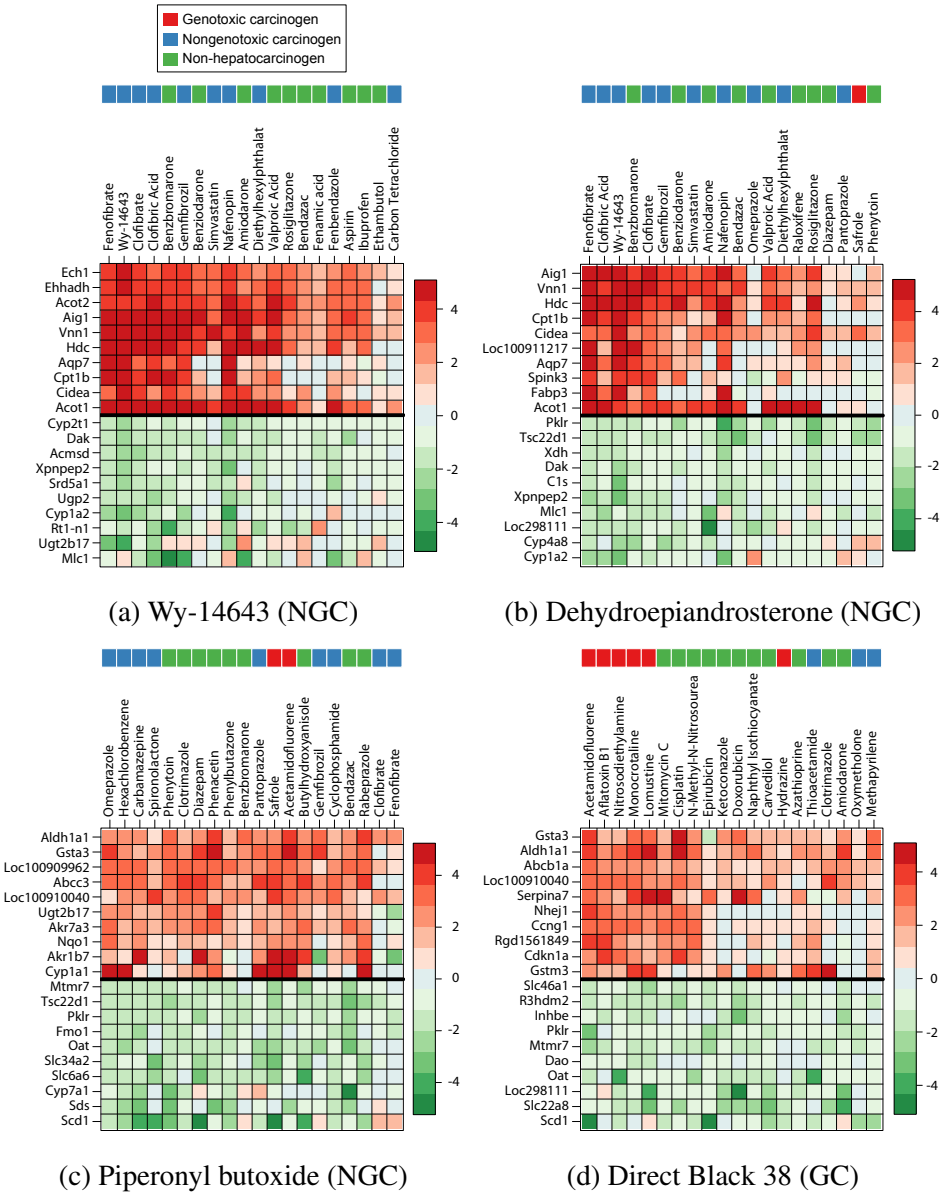
5.3.2 Identification of similar conditions

As the second step of our evaluation, we tested if we can use gene expression fingerprints and our similarity score to identify TG-GATEs and DrugMatrix experiments that use the same compound as experiments from an independent dataset. To this end, we extracted gene expression fingerprints for each compound in the evaluation dataset (see Section 5.1.5 and Table A.2). As the dynamic range of intensity ratios was very similar to those observed for TG-GATEs and DrugMatrix, we used the method described above to extract the fingerprints in the validation dataset. We collected genotoxicity and carcinogenicity information for the evaluation compounds from the CPDB where available. For the remaining compounds, we used annotations from the original publications (Ellinger-Ziegelbauer *et al.*, 2008; Römer *et al.*, 2014a). Then, we calculated the similarity coefficient S for all fingerprints of the validation dataset and the TG-GATEs and DrugMatrix fingerprints and ranked the TG-GATEs and DrugMatrix conditions by similarity.

Ten of the 15 compounds in the validation dataset are also contained in either TG-GATEs, DrugMatrix or both. Of these 10 substances, 8 are NGCs, one is an NC (nifedipine, NIF), and one is a GC (nitrosodimethylamine, DMN). For five of the eight NGCs (acetamide (AAA), ethionine (ET), methapyrilene (MP), phenobarbital (PB) and thioacetamide (TAA)) the condition that was most similar, i.e., had the highest similarity score S, was a TG-GATEs or DrugMatrix experiment using the same substance. The remaining three NGCs (cyproterone acetate (CPA), diethylstilbestrol (DES) and Wy-14643 (WY)) were only placed second in the respective rankings, but the most similar database condition was an experiment with a different NGC that has a very similar MOA. This indicates that the combination of the gene expression fingerprints with our similarity score can recall experiments with the same or similar compounds from TG-GATEs and DrugMatrix.

We found that the ranking of database compounds can provide leads for the MOA analysis for new compounds. For example, WY is an NGC that acts as a peroxisome proliferator-activated receptor alpha (PPARa) agonist and is investigated as a potential drug for treating cardiac dysfunction (Wölkart *et al.*, 2012). The database conditions that are most similar are experiments using fenofibrate, clofibric acid, and clofibrate (see Fig. 5.2(a)). These three NGCs are also known to interact with PPARa (Peraza, 2005), which suggests a PPARa-related MOA for WY (as shown by Peraza, 2005). For dehydroepiandrosterone (DHEA), an NGC that is not included in TG-GATEs and DrugMatrix, many known PPARa regulators are among the most similar database conditions: fenofibrate, clofibric acid, WY, and clofibrate (see Fig. 5.2(b)). Again, this indicates a PPARa-related MOA for DHEA, as was previously shown by Mastrocola *et al.* (2003).

For piperonyl butoxide (PBO), an NGC that is used as a component in pesticides, the most similar database conditions are experiments with the NGCs omeprazole, hexachlorobenzene, carbamazepine, and spironolactone (see Fig. 5.2(c)). Of these chemicals, omeprazole, hexachlorobenzene, and carbamazepine are considered enzyme inducers (Hayashi *et al.*, 2012; Uehara *et al.*, 2011), which suggests a similar, enzyme-



(a) Wy-14643 (NGC)

(b) Dehydroepiandrosterone (NGC)

(c) Piperonyl butoxide (NGC)

(d) Direct Black 38 (GC)

Figure 5.2: **Gene expression heat maps of similar compounds.** For selected test chemicals, we extracted the most similar chemicals included in either TG-GATEs or DrugMatrix. Each column corresponds to a chemical that was identified as similar. The chemicals are sorted from left to right by descending similarity score. The heat maps show the $\log_2$ fold change of 20 selected genes from the gene fingerprints of the test chemical. Genes above the black line are upregulated at least 1.5-fold in the test chemical, and genes below the line are downregulated. Genes were selected based on average expression in the identified chemicals. The color bar above the chemical name indicates the hepatocarcinogenicity annotation (legend is shown in (a)).

inducing MOA for PBO, as was demonstrated by Goldstein *et al.* (1973). For other enzyme inducers in the validation dataset, e.g., cyproterone acetate (CPR) and PB, we observed similar effects. Omeprazole, spironolactone, and carbamazepine are among the most similar compounds for CPR, suggesting enzyme induction as the major mechanism of carcinogenicity, as Schulte-Hermann *et al.* (1980) have demonstrated. For PB, carbamazepine and hexachlorobenzene are among the most similar compounds, which again suggests enzyme induction, as has been shown by Waxman *et al.* (1983). Sulfasalazine, which is classified as an enzyme-inducing NGC by Uehara *et al.* (2011), is also among the compounds most similar to PB, but because it has no positive test for hepatocarcinogenicity in the CPDB, we here considered it to be an NC.

For the validation compounds TAA, MP, and ET (all NGCs) that are considered hepatotoxic oxidative stressors according to Uehara *et al.* (2011), we also observed that compounds with similar MOA were found, although the evidence is weaker compared to the compounds described above. For TAA, the most similar compound is MP, but with a low similarity score S in comparison with the TAA experiments contained in TG-GATEs, which suggests some differences in the observed gene expression. Among the compounds most similar to MP are carbon tetrachloride and TAA, which supports the hepatotoxic MOA, but also the PPARa-activator gemfibrozil, and the genotoxic compound hydrazine. For ET, the most similar compounds include TAA and MP, as well as carbon tetrachloride, which is also a hepatotoxic oxidative stressor (Uehara *et al.*, 2011).

The genotoxic compound DMN was not recalled, which may also be due to the different dose level and duration in the experiments (10 mg/kg/day for five days in DrugMatrix vs. 4 mg/kg/day for seven days in the evaluation dataset). However, nitrosoethylamine (DEN), which is very similar to DMN chemically, was identified as the most similar compound for DMN, along with other GCs. Similarly, for Direct Black 38 (CIDB), the second GC, the five most similar compounds are all GCs, the highest similarity score was observed for acetamidofluorene (see Fig. 5.2(d)).

In summary, this evaluation demonstrates that database compounds with similar MOA are found when using our fingerprint similarity scoring and compounds that are contained in the validation dataset and one or both of the databases are recalled. This shows that the proposed similarity scoring can be used for mechanistic analysis, e.g., to identify leads for a MOA analysis or carcinogenicity evaluation.

5.3.3 Intensity ratio threshold evaluation

Next, we evaluated the thresholds that were used to extract the gene expression fingerprints. These were selected in accordance with other publications that commonly use 1.5- or 2-fold up- or downregulation tresholds to select DEGs. To validate our choice, we determined how the compounds that were found to be similar to a query substance share it's toxicological class. If a substance was present in the validation dataset and in either of the databases, the database experiments with this compound were not considered to ensure an unbiased evaluation.

We used the carcinogenicity and genotoxicity annotations from the CPDB to annotate the TG-GATEs and DrugMatrix conditions as described above and also annotated the validation dataset (see Table A.2). In detail, we counted the compounds that shared the class of the query compound among the 5, 10, and 20 most similar compounds (see Table 5.1). We observed that the lower threshold (1.5-fold deregulation) performed slightly better than the higher threshold. For the lower threshold, on average 4.3 (4.1 for the higher threshold) out of the 5, 8.0 (7.3) out of the 10, and 14.2 (14.0) out of the 20 compounds identified as the most similar shared same carcinogenicity class.

We also calculated relative similarity scores $\tilde{S}$ by dividing the observed similarity score for each query compound with the highest observed similarity score. We determined the percentage of conditions annotated with the same carcinogenicity class in the subset of conditions with a relative similarity score higher than 0.8 and 0.7 (see Table 5.1). Again, we observed a slightly better performance for the less conservative fold change cutoff. On average, 88% and 80% of the identified conditions with a $\tilde{S} > 0.8$ were of the same class as the evaluation chemical, while only 80% and 78% conditions with matching classes were found for $\tilde{S} > 0.7$.

In conclusion, our evaluation shows that our similarity scoring approach can robustly identify compounds with similar genotoxic and hepatocarcinogenic potential. The identification is possible for chemicals that are present in one or both databases as well as for compounds that are not included in any of the two databases. Across all evaluations, the 1.5-fold deregulation threshold led to better results for the similarity search. One possible explanation may be the larger number of genes available for the similarity scoring. For the lower threshold, the median fingerprint size is 269 genes, compared to 53 genes for the higher threshold. We found that the number of differentially regulated genes was particularly small for NGCs and NCs, which may lead to problems when distinguishing these two substance classes. Based on the above evaluations, we found that the 1.5-fold deregulation threshold performed better and consider conditions with a relative similarity score $\tilde{S} > 0.8$ as likely to share the same class.

Table 5.1: **Percentage of correctly identified conditions.** The most similar conditions were extracted for each chemical in the evaluation set. The percentage of conditions with the same carcinogenicity class in the 5, 10 and 20 most similar conditions and for conditions with a relative similarity $\tilde{S}$ above 0.8 and 0.7 was calculated.

Neighborhood	1.5-Fold Deregulation	2-Fold Deregulation
Top 5	86	82
Top 10	80	73
Top 20	71	70
$\tilde{S} \geq 0.8$	88	80
$\tilde{S} \geq 0.7$	80	78

5.3.4 Hepatocarcinogenicity prediction

Above, we demonstrated that gene expression fingerprints and our similarity score can be used to find leads for mechanistic analysis and identify compounds with similar toxicological class. We proposed the use of a relative similarity $\tilde{S}$ above 0.8 as a threshold for identifying compounds that are likely to have similar MOAs and toxicological properties. As the last step of our validation, we tested if we can use the extracted conditions to predict the toxicological properties, here hepatocarcinogenicity and genotoxicity of new compounds.

We used the lower threshold (1.5-fold up- or downregulation) because the number of identified compounds with the same toxicological class was highest (see Table 5.1). Again, we compared the gene expression fingerprints of the validation compounds with all database conditions and calculated the relative similarity scores. We consider all conditions with a relative similarity $\tilde{S} \geq 0.8$ to be similar, whereas conditions that do not meet this threshold are considered different. Then, we calculated the percentage of GCs (R_{GC}) and NGCs (R_{NGC}) in the conditions identified as similar and used an over-representation test to assess if there are more GCs or NGCs in the set of similar conditions than would be expected by chance. To estimate the probability of observing the actual or a higher ratio by chance, we performed a random permutation test with $n = 100,000$ repetitions. In each repetition, randomly drawn gene expression fingerprints were scored against the database to estimate the distribution of the R_{GC} and R_{NGC}. This results in a p-value, which is calculated as $p_{\mathrm{GC}} = \frac{N}{n}$, where N is the number of random gene fingerprints that contained a higher ratio of GCs. Analogously, p_{NGC} was computed. Using these p-values and a false discovery rate threshold of 0.05, we classified a validation compound as a GC if $p_{\mathrm{GC}} < 0.05$, or as an NGC if $p_{\mathrm{NGC}} < 0.05$, or as an NC if $p_{\mathrm{GC}} > 0.05$ and $p_{\mathrm{NGC}} > 0.05$. As the results in Table 5.2 show, the true class was predicted for all 15 validation chemicals. This shows that we can predict toxicological properties using gene expression fingerprints, the relative similarity of these fingerprints, and a random permutation-based over-representation test.

5.4 The ToxDBScan web application

We have implemented the similarity scoring of the TG-GATEs and DrugMatrix database in ToxDBScan, a web application which provides a simple interface that life science researchers can use to submit a query fingerprint. To perform a query, the researcher submits only the list of up- and downregulated genes in separate input boxes and selects the type of identifier that was used by the researcher. No confidential data needs to be uploaded, such as the chemical structure, compound name, or experimental details. Currently, ToxDBScan supports official rat gene symbols (as provided by the Rat Genome Database (Laulederkind *et al.*, 2013)), Entrez IDs (Maglott, 2004), Ensembl IDs (Flicek *et al.*, 2014), and UniProt IDs (The UniProt Consortium, 2014). The user can also choose

Table 5.2: **Classification results.** Similar conditions in TG-GATEs and DrugMatrix were identified by computing the similarity score S and selecting conditions with a relative similarity $\tilde{S} > 0.8$. Ratios of genotoxic carcinogens (R_{GC}) and nongenotoxic carcinogens (R_{NGC}) were calculated based on the annotation of the similar conditions. A permutation test ($n = 100,000$) was performed to assess the significance of over-representation of GCs (p_{GC}) and NGCs (p_{NGC}). If the p-values were significant for $\alpha = 0.05$, the corresponding class was predicted. If no significant enrichment was found for either of the two classes, the test chemical was predicted to be an NC. Significant p-values are printed in bold font.

Chemical	R_{GC}	p_{GC}	R_{NGC}	p_{NGC}	**Prediction**
Genotoxic carcinogens					
CIDB	1.00	**0.001**	0.00	1.000	GC
DMN	0.70	**0.008**	0.20	0.615	GC
Nongenotoxic carcinogens					
MP	0.00	1.000	1.00	**0.007**	NGC
TAA	0.00	1.000	1.00	**0.012**	NGC
DES	0.00	1.000	1.00	$<$ **0.001**	NGC
WY	0.00	1.000	1.00	$<$ **0.001**	NGC
PBO	0.00	1.000	0.77	**0.002**	NGC
MCA	0.00	1.000	1.00	**0.017**	NGC
AAA	0.00	1.000	0.60	**0.048**	NGC
DHEA	0.00	1.000	0.88	$<$ **0.001**	NGC
ET	0.00	1.000	1.00	$<$ **0.001**	NGC
CPR	0.00	1.000	0.71	**0.018**	NGC
PB	0.00	1.000	1.00	**0.019**	NGC
Non-hepatocarcinogens					
CFX	0.00	1.000	0.00	1.000	NC
NIF	0.25	0.132	0.25	0.399	NC

between the lower and the higher threshold to account for the strictness that was used to generate the query fingerprint. Additional options include the choice to use only one of the two databases or exclude compounds without carcinogenicity annotation from the database search. The web application ToxDBScan is publicly available from the ZBIT Bioinformatics Toolbox (Römer *et al.*, 2016b), which is described in detail in Chapter 4.

Similarity search report

ToxDBScan writes an HTML report that contains all information on the results of similarity search in TG-GATEs and DrugMatrix (see Fig. 5.3). The HTML report is displayed directly inside the ZBIT Bioinformatics Toolbox. Researchers can download the HTML report as a compressed file and view the HTML file in any modern browser. This report includes the results of the database scan for similar compounds and information on the NGC-specificity and information content of the differentially regulated genes in the

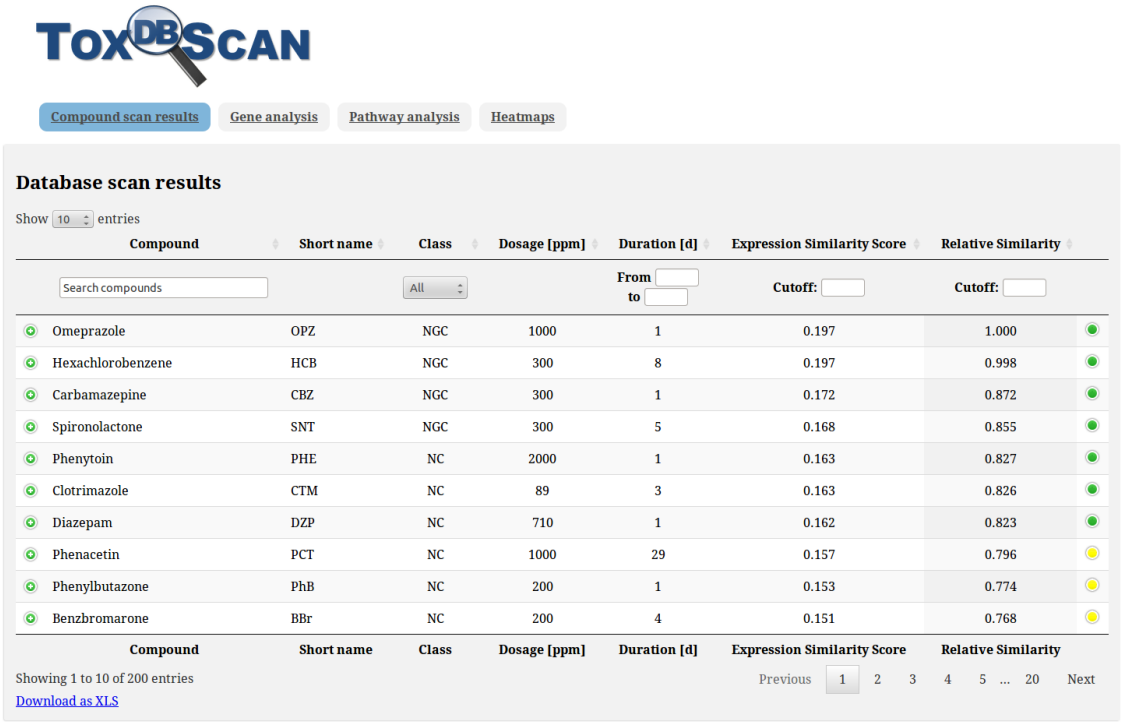


Figure 5.3: **HTML report of the compound similarity scan for PBO.** This figure shows the results of the similarity search against TG-GATEs and DrugMatrix. Additional information for each compound can be shown by clicking on the "plus" in the first column of the table. Additional information on the deregulated genes is available from the "Gene analysis" tab at the head of the report. The results of the pathway enrichment analysis against the KEGG database are available from the "Pathway analysis" tab. The "Heatmaps" tab shows heat maps of the gene expression in the most similar compounds.

query fingerprint. The results of the database scan include the similarity score and relative similarity score, the gene expression fingerprint for each condition, statistics on the genes that are shared between fingerprints, and links to external information on genes and compounds, e.g., CAS number, structure, or Entrez gene entries. For each condition, the CPDB annotation for the administered compound is included. This information can be used for a mechanistic analysis of the hepatocarcinogenic potential or MOA detection. The report also includes heat maps that show the gene expression of the most similar conditions. All tables can be downloaded as a PDF for printing and sharing or in tabular format for further analysis.

Pathway enrichment report

In addition to the database similarity search, ToxDBScan performs a pathway enrichment analysis with KEGG pathways (Kanehisa *et al.*, 2014). To this end, gene symbols were mapped to the corresponding *Rattus norvegicus* pathways, which we obtained from the KEGG database. For each pathway in the KEGG database, a hypergeometric test was performed to check for significant pathway perturbation. The enrichment p-value is computed as:

$$P(X \geq m) = \sum_{i=m}^{M} \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} \tag{5.8}$$

where N is the number of all genes for which gene expression was measured, M is the number of genes in the pathway of interest, n is the number of DEGs and m is the number of DEGs that are part of the pathway of interest. The resulting p-values were corrected for multiple hypothesis testing with Benjamini–Hochberg correction (Benjamini and Hochberg, 1995). The results of the pathway enrichment analysis are available from a separate tab in the HTML report. For each pathway, this pathway enrichment tab provides the p-value, information on the genes that are part of the pathway, a pathway map from the KEGG database, and links to external resources, e.g., the corresponding web page in the KEGG database.

5.5 Summary and conclusions

In this chapter, we presented a similarity coefficient for gene expression fingerprints, which can be used to assess the similarity of gene expression profiles that were collected in transcriptomics studies. We applied our fingerprint extraction method to gene expression data from TG-GATEs and DrugMatrix and evaluated our method for MOA discovery, identification of compound with similar toxicological properties, and predictive analysis of hepatocarcinogenicity. Using a dataset of microarray experiments that are not part of either TG-GATEs and DrugMatrix, we demonstrated that our new method can be used to robustly identify toxicologically similar compounds and predict hepa-

tocarcinogenicity and genotoxicity. This evaluation dataset encompasses 15 chemicals, which are either GCs, NGCs, or NCs.

The three major MOAs for the evaluation NGCs were oxidative stress-mediated hepatotoxicity (TAA, MP, ET), PPARa-regulation (WY, DHEA), and enzyme induction (PB, PBO, CPR) (Peraza, 2005; Uehara *et al.*, 2011; Schulte-Hermann *et al.*, 1980; Goldstein *et al.*, 1973; Mastrocola *et al.*, 2003). We demonstrated that the database conditions that were identified to be the most similar induce toxicity through the same MOA. Furthermore, GCs in the databases were identified as most similar to the genotoxic evaluation chemicals (CIDB, DMN). This demonstrates that the proposed similarity scoring approach using gene expression fingerprints provides a useful tool for researchers that want to use TG-GATEs and DrugMatrix to investigate the MOA of new compounds or drug candidates.

We explored several strategies to establish a neighborhood of similar compounds based on the ranking that is produced by the database similarity search. The best results were obtained with a low intensity ratio threshold (1.5-fold up- or downregulation) and a relative similarity cutoff that only considers compounds as similar if they have a relative similarity $\tilde{S} > 0.8$. Using these parameters, we observed that 88% of the TG-GATEs and DrugMatrix conditions that were found to be similar share the carcinogenicity class of the evaluation compounds. Subsequently, we performed a predictive analysis in which we determined the ratio of genotoxic and carcinogenic substances in the similarity neighborhood and compared this ratio to a background ratio that was estimated using random permutation sampling. We were able to correctly predict all 15 evaluation chemicals as NGC, GC, or NC, respectively. This indicates that the proposed method can be used to predict the hepatocarcinogenic potential of new compounds based on these two large databases of compounds with known hepatocarcinogenic potential.

ToxDBScan is freely available as part of the ZBIT Bioinformatics Toolbox and was developed to enable other researchers to use our similarity score for the identification of similar compounds in TG-GATEs and DrugMatrix. ToxDBScan is independent of the platform used to identify the deregulated genes, as only the list of up- and down-regulated genes is required to run ToxDBScan. At the time of writing, we are not aware of other web applications which offer a similarity search in both TG-GATEs and DrugMatrix. With LTMap by Xing *et al.* (2014), a similar web application exists for scanning the TG-GATEs database. However, LTMap does not scan DrugMatrix, nor provide the additional analysis performed by ToxDBScan, e.g., pathway enrichment analysis or calculation of the information content of genes.

In conclusion, our similarity scoring approach for gene expression profiles and the new tool ToxDBScan offer a novel and unique similarity scoring method for the two largest toxicogenomics databases and may contribute to the implementation of new approaches to the evaluation of the carcinogenic potential of chemicals.

Chapter 6

Multi-*omics* approaches for prediction of nongenotoxic carcinogenicity

The previous chapter presented ToxDBScan, a web tool that can be used to query TG-GATEs and DrugMatrix for compounds that induce gene expression changes similar to those observed for new compounds. This could help identify potential side effects and characterize the modes of action of new drugs. We also demonstrated that ToxDBScan can be used to predict the carcinogenic potential of new drugs using microarray data and thus help replace the lifetime rodent cancer bioassay (LRB) with short-term alternatives.

Several other groups have also developed systems or methods for the identification of nongenotoxic compounds in short-term assays, both *in vivo* and *in vitro*. Similar to ToxDBScan, most of these groups used a combination of mRNA expression profiling and machine learning algorithms or statistical methods to perform the prediction (see for example the reviews by Waters *et al.* (2010) and Afshari *et al.* (2011)). Auerbach *et al.* (2010) proposed the use of these short-term prediction systems to prioritize environmental or industrial chemicals for long-term carcinogenicity bioassays. Other studies showed that toxicologists can gain mechanistic insights from these toxicogenomics studies, e.g., specific molecular profiles can be associated with toxicological phenotypes and adverse effects observed in animal studies (Afshari *et al.*, 2011). Among others, Ellinger-Ziegelbauer *et al.* (2005) published a signature, i.e., a list of early biomarker genes, for discriminating GCs and NGCs in Wistar-Hannover rats after short (up to two weeks), repeated administration. They report that their signature reflects established GC and NGC modes of action: for GCs, a strong DNA damage response is observed, whereas an increased cell cycle progression is the dominant feature for NGCs (Ellinger-Ziegelbauer *et al.*, 2005). In a follow-up study, they collected mRNA expression data for a larger set of compounds and trained a prediction model using SVMs (Ellinger-Ziegelbauer *et al.*, 2008). Similarly, Japanese researchers published two signatures for discriminating NGCs and NCs based on mRNA expression data collected in Sprague-Dawley rats (Uehara *et al.*, 2008, 2011). In their first study, Uehara *et al.* (2008) used an experimental setup with a single drug exposure and collected liver samples after 24 hours. In the second study, Uehara *et al.* (2011) used a repeated dosing setup, in which animals received a daily administration of one compound for 28 days, similar to the study by Ellinger-

Ziegelbauer *et al.* (2008). The data used in these studies has been made available in the TG-GATEs database (Uehara *et al.*, 2010). The results that were observed by Uehara *et al.* (2011) and Auerbach *et al.* (2010) suggest that longer dosing may allow a better detection of carcinogenic activity, which needs to be weighed against the desired early detection of NGCs. Two groups also explored the use of different *omics* layers: Schmitz-Spanke and Rettenmeier (2011) used a protein expression profiling approach and Yokoi and Nakajima (2011) explored the use of microRNAs (miRNAs) for toxicogenomics.

The development of short-term assays that can reliably predict NGCs and thus help prioritize compounds that are most likely NCs for development would reduce the investment of animals, time, and money in chemical and drug development. Currently, most groups used mRNA data and only a few groups explored other *omics* layers, but no group combined multiple *omics* layers to build an integrated prediction system. Tong *et al.* (2009) proposed the integration of multiple regulatory layers for the analysis of NGC mechanisms, which was later repeated by Khan *et al.* (2014). Here, we propose a new, holistic approach that uses multiple *omics* layers to predict the hepatocarcinogenicity of drug candidates or other chemical substances. Building on the previous research with single *omics* approaches, we introduce two new concepts: the integration of *omics* data for multiple regulatory and functional biological layers (mRNA, miRNA, and protein expression) and the abstraction from individual signature genes to higher-order levels, such as pathway enrichments or molecular interactions. We demonstrate the value of these new concepts with a unique data resource that combines mRNA expression profiles with miRNA and protein expression profiles obtained from male Wistar rats for multiple GCs, NGCs, and NCs after up to 14 days of daily compound administration. We use CV to show that the predictive power of mRNA signatures can be increased by adding complementary *omics*-based features obtained from profiling other molecular levels. Furthermore, we demonstrate that an additional improvement can be gained by including complex, integrative features. The content of this chapter was published in *PLoS ONE* under the title "Cross-platform toxicogenomics for the prediction of nongenotoxic hepatocarcinogenesis in rat" (Römer *et al.*, 2014a).

6.1 New integrative feature representations for multi-*omics* data

As pointed out in Chapter 2, gene regulation is a complex process that is a result of many layers of regulation. These layers of regulation interact not only with the DNA locus of a gene but also with the intermediate products of transcription and translation. For example, miRNAs bind to complementary base sequences in mRNAs and thus silence the mRNA by inhibiting the translation of the mRNA to proteins (Bartel, 2009). For this reason, the abundance of mRNAs that is translated into a certain protein may not correlate perfectly with the amount of protein that is present in a cell. The direct

assessment of the protein abundances in a cell provides the only reliable account of the actual protein abundance. However, this direct assessment is currently not possible on a proteome-wide scale because quantitative mass spectrometry is not yet mature enough for reliable, large-scale application and targeted, antibody-based RPPAs capture only a small, preselected fraction of all proteins in a cell. By taking additional layers of regulation into account, the uncertainty that is associated with traditional mRNA expression profiling may be reduced, which in turn may provide better predictors for biological and toxicological questions. To achieve this, we developed two new concepts that are designed specifically for multi-*omics* data and provide an integrated view of gene regulation: molecular interaction features, which capture the interaction of molecules from two *omics* layers, and pathway enrichment features, which provide an additional level of abstraction and incorporate information from external pathway databases.

6.1.1 Molecular interaction features

The molecular interaction features integrate the data observed for two interacting molecules, which were measured by two different *omics* technologies into a single interaction score (see Fig. 6.1A). To account for the varying dynamic ranges of differential expression for different *omics* technologies, we first transform the observed log-ratios, i.e., $\log_2$(fold changes), to a common scale. This transformation is performed for each technology and assures that each *omics* platform contributes equally to the final interaction score, independent of the dynamic range. All log-ratios are transformed using a simple linear scaling function $s : x \rightarrow [-1,1]$ with

$$s(x) = \frac{x}{max(-min(\mathbf{F}), max(\mathbf{F}))}. \tag{6.1}$$

where $\mathbf{F} \in \{\mathbf{F}_{mRNA}, \mathbf{F}_{miRNA}, \mathbf{F}_{protein}\}$ is the set of all log-ratios observed for the source *omics* platform. The molecular interaction score $m(x_i, x_j)$ for each potential interaction between two molecules (e.g., miRNA i and target mRNA j) is calculated by the linear combination of the observed log-ratios for each molecule, i.e., the product of the scaled log-ratios:

$$m(x_i, x_j) = s(x_i) \cdot s(x_j) \tag{6.2}$$

We expect $m(x_i, x_j)$ to be close to 1 if the expression of the molecules i and j is positively correlated (e.g., mRNA and protein product) and close to -1 for inhibiting interactions (e.g., miRNA and target mRNAs).

To identify interacting molecules, we used external databases and biological principles. We gathered potential interactions between miRNAs and mRNAs from curated databases of validated miRNA targets (TarBase v5.0c (Papadopoulos *et al.*, 2009), miRTarBase v2.4 (Hsu *et al.*, 2011), and miRecords v3 (Xiao *et al.*, 2009)) as well as miRNA target prediction tools (ElMMo v5 (Gaidatzis *et al.*, 2007), DIANA-microT v4.0 (Maragkakis *et al.*, 2009), and TargetScan v5.2 (Lewis *et al.*, 2003)). An mRNA and a

protein were considered to interact if the protein is the product of the mRNA's translation. We also defined interactions between miRNAs and proteins such that a miRNA is considered to interact with a protein if the protein is translated from an mRNA that is targeted by the miRNA. We expect to find mainly negative correlations between miRNAs and their targeted mRNAs (because miRNAs inhibit translation of target mRNAs), positive correlations between mRNAs and translated proteins, and negative correlations between miRNAs and proteins that are translated from the targeted mRNAs.

6.1.2 Pathway enrichment features

The pathway enrichment features integrate data for several data points from multiple *omics* platforms into a single pathway score. This is achieved through an abstraction from genes (or other genetic molecules such as miRNAs or protein) to pathways (see Fig. 6.1B). This abstraction could provide a more robust representation of a compound's toxicological effects in an organism because changes in different genes or other regulatory molecules may contribute to an alteration of the same pathway and thus to similar effects.

To integrate the observed effects across multiple *omics* platforms, we combine the lists of molecules that are found to be differentially expressed in each platform into a common list. For this purpose, all profiled mRNAs and proteins were mapped to their corresponding genes, which can, in turn, be attributed to canonical pathways. The integration of miRNAs into this process is more difficult because most canonical pathway databases do not include miRNAs in their pathways. For this reason, we modeled the effect of deregulated miRNAs through the targeted mRNAs. Next, we used a hypergeometric test to calculate p-values for pathways, which we had previously obtained from pathway databases. We do not include topological information from the network of the pathway. The p-values are calculated with the following formula:

$$p = \mathrm{P}(X \geq m) = \sum_{i=m}^{M} \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} \tag{6.3}$$

where N is the number of all genes that were measured, M is the size of the pathway, n is the size of the combined list of differentially regulated genes on all *omics* platforms, and m is the number of the genes that are also in the pathway. The final pathway enrichment score is calculated as $-log_{10}(p)$. Thus, the higher the enrichment score computed for a certain pathway, the more significant is the overrepresentation of genes from the combined list in this pathway. The hypergeometric test is performed for each pathway to construct the pathway enrichment feature vector for a compound. This feature vector represents both an abstraction from the level of single genes to pathways and an integration of multiple *omics* platforms. We used three canonical pathway databases to calculate pathway enrichment scores: KEGG (Kanehisa and Goto, 2000), Reactome (Matthews *et al.*, 2009), and BioCarta (Nishimura, 2001).

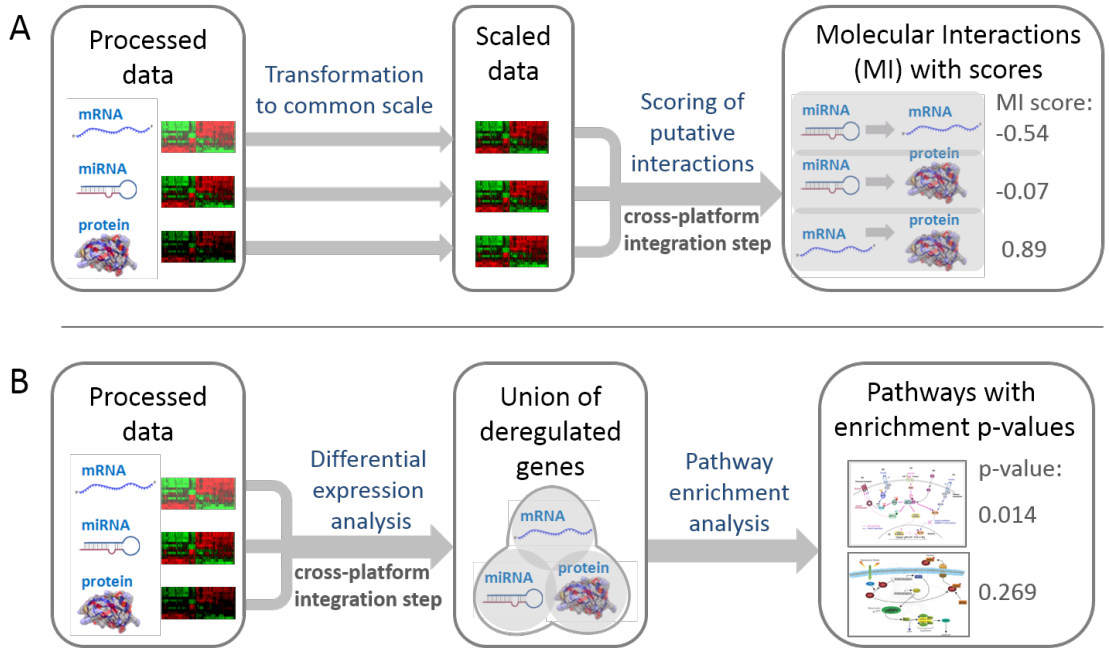


Figure 6.1: **Feature representations used for cross-toxicogenomics prediction models.** (A) Molecular interaction features. The processed data from the different platforms, given in the form of $\log_2$-transformed fold changes, were mapped to the same interval (here: [-1, 1]) using a linear function in order to account for the different dynamic ranges of the platforms. Next, putative interactions between molecules represented on different platforms were inferred based on negatively or positively correlated expression profiles. For miRNAs, all possible interactions with experimentally validated and predicted mRNA targets were considered. Associations between mRNAs and proteins were made based on common gene loci. The connections between miRNAs and proteins can be transitively inferred from the corresponding mRNA interactions. In order to obtain a numeric feature representation, a score was computed for each interaction, which equals the product of the scaled log-ratios calculated for the two interacting molecules. (B) Pathway enrichment features. First, differentially expressed features were detected for each platform separately based on appropriate fold change and/or p-value cutoffs. All transcripts and proteins were mapped to the corresponding genes in order to facilitate their association with metabolic and signaling pathways. As miRNAs are typically not contained in canonical pathways, deregulated miRNAs were represented by the genes corresponding to their experimentally confirmed target mRNAs in order to model their impact on pathways. The union of deregulated genes was computed across platforms. Then a hypergeometric test was applied to determine enriched pathways represented by these genes. Finally, a feature vector was constructed, representing the $\log_{10}$-transformed p-values obtained for each pathway from the overrepresentation test. Figure from Römer *et al.* (2014a).

6.2 Model construction workflow with single- and multi-*omics* features

The study that we describe in this chapter was designed to explore the potential of integrated molecular signatures. In total, we collected data from four molecular profiling techniques: traditional mRNA microarrays, miRNA microarrays, DNA methylation arrays and RPPAs. Some of these techniques are not yet as mature as the established mRNA microarrays, which leads to higher costs and less reliability. For this reason, the number of substances was small in comparison with other toxicogenomics studies. We also had to reject the data collected with DNA methylation arrays to due experimental problems, which led to a significant bias in the observed methylation patterns. In total, we have collected data with three *omics* technologies for 15 compounds, of which 8 are NGCs, 2 are GCs, and 2 are NCs (see Table A.2). For each compound, three biological replicates were profiled. We excluded three of the study compounds (CPA, TAA, and WY) from the model construction process because the classification of these was uncertain. To reach a sample size that allows the application of machine learning, we decided to use the data from each biological replicate as a sample instead of summarizing the data of the replicates. This results in a dataset of 36 samples for training and evaluation. Due to the small number of samples, we chose a nested CV procedure over an external validation set. The following sections describe the data processing and model construction in more detail.

6.2.1 Data preprocessing

The tissue samples for this study were collected from the liver of 8 to 10 weeks old, male Wistar Hanover rats (strain Crl:WI[Gl/BRL/Han]IGS BR), which were assigned to treatment groups using a weight stratification-based computer program. Each group of animals received either a daily dose of one substance or solvent control for up to 14 days. The time points at which the tissues for this study were collected are listed in Table A.2. The administered doses were selected based on those reported to induce liver tumors in the LRB (Ellinger-Ziegelbauer *et al.*, 2008). We used time-matched groups that were treated with the corresponding vehicles (methylcellulose (MC) or corn oil (CO)) to serve as a reference for the calculation of expression changes for all *omics* molecules.

Total mRNA expression was profiled with Affymetrix GeneChip RAE230A microarrays and data was exported as CEL files. The RAE230A array contains 15,866 probe sets, which correspond to 5,399 annotated rat genes and 10,467 expressed-sequence tags. We used the package `arrayQualityMetrics` (Kauffmann *et al.*, 2009) for R/Bioconductor (Gentleman *et al.*, 2004) to assess the quality and validate that no experimental problems were present. Next, we used the package `affy` (Gautier *et al.*, 2004) to perform background correction, normalization between arrays, and probe summarization with RMA normalization.

The Agilent Rat miRNA Microarray 1.0 (G4473A) was used to profile miRNA expression. Again, sample quality was assessed with the `arrayQualityMetrics` package. We performed the normalization with a variant of the RMA algorithm as proposed by López-Romero (2011). To identify the putative molecular interactions between miRNAs and mRNAs, we collected experimentally confirmed and predicted target mRNAs for all miRNAs. We obtained information for validated mRNA-miRNA interactions from TarBase v5.0c (Papadopoulos *et al.*, 2009), miRTarBase v2.4 (Hsu *et al.*, 2011), and miRecords v3 (Xiao *et al.*, 2009) and predicted interactions with the miRNA target prediction tools ElMMo v5 (Gaidatzis *et al.*, 2007), DIANA-microT v4.0 (Maragkakis *et al.*, 2009), and TargetScan v5.2 (Lewis *et al.*, 2003).

Protein expression was profiled with ZeptoMARK RPPAs as described by (Pirnia *et al.*, 2009). In short, specific primary antibodies were used to perform a two-step immunoassay and detect proteins and protein modifications. The measured signal intensity was background corrected and quantile normalized within each array. Missing values, which accounted for approx. 1% of the data, were estimated using *k*-Nearest-Neighbor imputation.

For all *omics* platforms, we calculated sample-wise intensity ratios by dividing the observed signal intensity for a molecule by the mean intensity that was observed for the time-matched vehicle control. Finally, the intensity ratios were $\log_2$-transformed to obtain the sample-wise, $\log_2$ fold changes, which we used as features for single-*omics* models and to calculate the integrative multi-*omics* features.

6.2.2 Inference of predictive molecular signatures

A predictive molecular signature is a list of biomarkers, e.g., mRNAs or proteins, which can be used to predict the toxicological properties of a compound. To filter predictive biomarkers from the thousands of measured biomolecules, we used recursive feature selection (RFE) with SVMs, which is called SVM-RFE. This method is well established and has been used by other toxicogenomics researchers before, e.g., by Ellinger-Ziegelbauer *et al.* (2008). The SVM-RFE method trains an SVM with a given set of features and uses the weight vector, which was optimized during SVM training, to rank the features by their relevance for the classification (Guyon *et al.*, 2002). This procedure is repeated on a reduced feature set in the next iteration after elimination of the least informative features. For example, the 10% least informative features are removed in each iteration, and this process is repeated until the desired signature size is reached or all features have been eliminated.

Then, we selected the best signature based on the predictive power of signatures of different sizes. We assessed the predictive power as the mean prediction accuracy in a nested CV with five different machine learning methods. More specifically, we assessed signatures containing 5, 10, 15, 20, and 25 features for their predictive power. Based on the mean accuracy that was observed for each of the signature sizes, we used spline interpolation to approximate the optimal signature size numerically.

To build a common signature across multiple repetitions of the CV evaluations, we merged the signatures which were obtained for each repetition into a rank-based consensus signature. For each signature, the rank of all features is determined by the SVM-RFE weights and each feature is ranked in the final signature by the average observed rank in the individual signatures. The consensus signature was then constructed by selecting the best features until the approximated optimal signature size was reached.

6.2.3 Validation of prediction models

Due to the small number of compounds in this exploratory study, we used 2×2-fold nested CV to evaluate the model performance (see Fig. 6.2). The nested CV scheme ensures that we perform an unbiased evaluation, as the parameter tuning is performed in the inner CV. The performance of each model is thus evaluated with independent samples, which were not seen during the model construction and tuning process. To minimize selection bias, which could lead to overestimation of the model performance, we repeated the CV 10 times with different random splits of training and validation data. Furthermore, to obtain an unbiased estimate of the prediction accuracy that can be achieved with the various single- and multi-*omics* features, we used five different classification methods to evaluate the predictive molecular signatures. These five classification methods are linear SVMs, random forests (RFs), Neural Networks (NN), Bayesian Generalized Linear Models (BGLM), and Principal Component Regression (PCR). RF, NN, BGLM, and PCR are implemented in the `caret` pacakge for R (Kuhn, 2008). The SVM was used via the R interface provided by the SHOGUN machine learning toolbox (Sonnenburg *et al.*, 2010). We used the AUC as the primary performance indicator and averaged the AUC obtained with the five classifiers to assess the predictive power of each single- or multi-*omics* signature.

6.3 Results of the model evaluation

As described above, we evaluated the performance of all models with a 2×2, nested CV, which we repeated 10 times to eliminate a potential selection bias. In the following section, we will report the results that we observed, as well as analyze the consensus signatures for the single- and multi-*omics* features and the classification of the three compounds that are not clearly assigned to a carcinogenicity class.

6.3.1 Classification performance of *omics* signatures

The primary goal of the application of toxicogenomics to compound carcinogenicity is the identification of lists of biomarkers, also called molecular signatures, which allow the prediction of tumor development before any histopathological changes can be observed. To this end, we inferred molecular signatures from features that were obtained

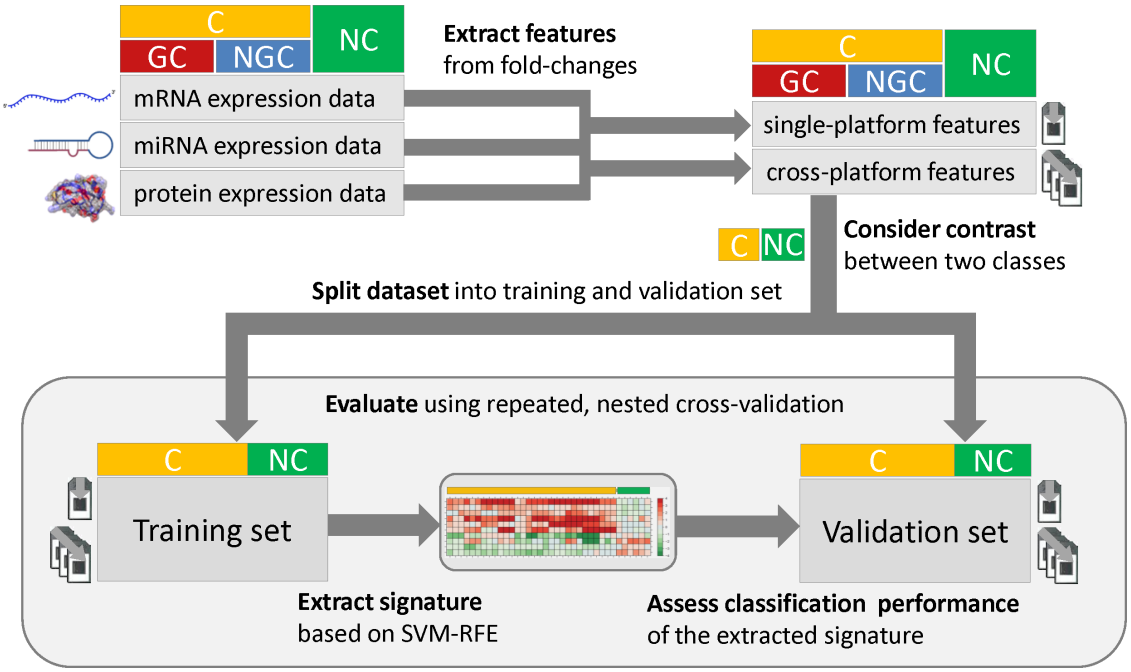


Figure 6.2: **Workflow used for signature extraction and evaluation of classification performance.** For the multi-level *omics* data available in this study, which includes mRNA, miRNA, and protein expression profiles of diverse compounds, fold changes were calculated for each gene and sample that could be confidently assigned to a certain compound class (C: carcinogens, GC: genotoxic carcinogens, NGC: nongenotoxic carcinogens, NC: noncarcinogens). Single-platform features simply correspond to fold changes observed on each specific biological level. In contrast, cross-platform features capture molecular interactions and pathway alterations, which can be inferred by integrating *omics* data across multiple layers. For each class contrast (e.g., C vs. NC) of interest, the dataset was split into a training set and a validation set. Using the SVM-RFE feature selection technique, a predictive signature for class discrimination was extracted, which was then used to predict the carcinogenic class of the samples in the validation set. By embedding this process into a 2-fold CV which was repeated 10 times with different random splits of the data, the classification performance can be robustly estimated based on the mean AUC. Figure from Römer *et al.* (2014a).

from three different *omics* platforms and from complex features that integrate the features from multiple *omics* platforms. We developed two types of complex features: molecular interaction (MI) features, which score putative molecular interactions based on miRNA-mRNA target and mRNA-protein relationships and pathway enrichment (PE) features, which encompass pathway level perturbations in the hepatic cells (Fig. 6.1). To assess if different *omics* platforms yield a different predictive power, we evaluated the signatures from each *omics* platform separately. We also build a combined signature, which was generated by merging the single-platform signatures, to determine if the profiling of multiple layers of gene expression improves the predictive power. This combined signature was further expanded into a hybrid signature by adding the signatures that were inferred with the MI and PE features to evaluate if the complex features improve the predictive power. To identify the contribution of each complex feature type, we evaluated the combination of the multi-*omics* signature with each complex feature type individually and with both complex feature types, which results in three different hybrid signatures: multi-*omics* and MI features, multi-*omics* and PE features, and multi-*omics* and both MI and PE features. This signature extraction process was performed for three different class contrasts (C:NGC+GC vs. NC, NGC vs. GC, NGC vs. NC) and evaluated with five supervised classification methods (Fig. 6.2). We determined the predictive power by calculating the average AUCs observed across the 10 repetitions for each contrast, classifier, and signature type (single-platform, combined, hybrid).

The results of the signature evaluation are shown in Fig. 6.3. In all three contrasts, we observed a higher mean AUC for the combined, multi-*omics* signature compared with the single-platform signatures. The only exception is the single-platform, protein signature in the NGC vs. GC contrast, which performed better than the combined signature (Fig. 6.3B). This indicates that the combination of biomarkers from multiple regulatory layers can increase the predictive performance, at least in this exploratory study with a small number of compounds. Moreover, we observed a better performance of the combined signature compared to the mRNA signatures, which are the current standard in most toxicogenomics studies. The miRNA signature performs worse than all other signatures in all cases, which indicates that miRNAs alone are not sufficient to reliably detect tumorigenic effects.

We also observed an increased AUC when adding the complex, integrative feature types to the combined signature. In two contrasts, C vs. NC and NGC vs. GC, the hybrid signatures performed consistently better than the multi-*omics* signature alone (Fig. 6.3A and C) and achieved higher mean AUCs. In the NGC vs. NC contrast, we observed only slight changes in the mean AUCs which does not allow a clear decision on an improvement. However, the performance of the hybrid signatures, particularly of the hybrid signature that encompasses both MI and PE features, was always at least as good or better than the combined signature and achieved the best AUCs in two of the three contrasts and the second best in the C vs. NC contrast. This indicates that the complex features can capture meaningful signals and thus add to the predictive power of the classifier. Furthermore, we observed a large variance in the AUCs of the single-platform signatures

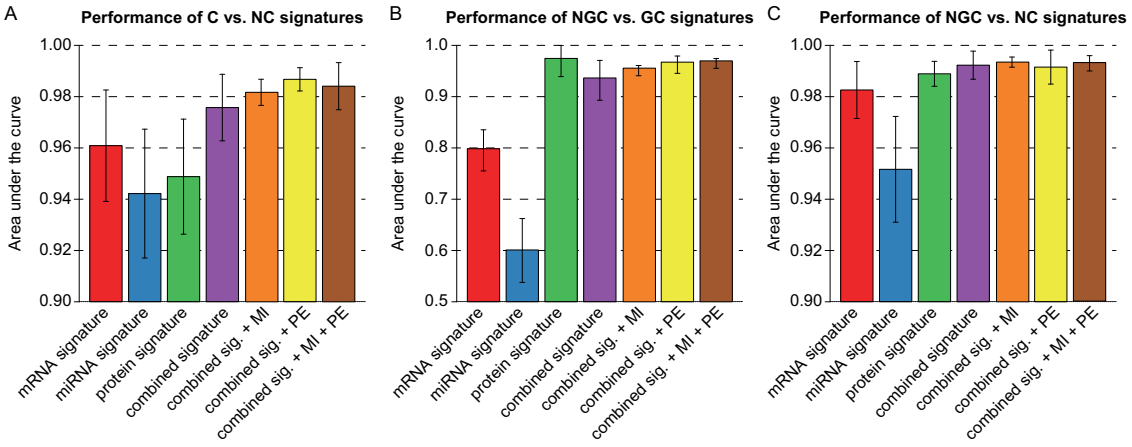


Figure 6.3: **Classification performance for different class contrasts depending on signature types.** The bar plots correspond to the average AUC obtained from five widely used supervised classification methods (SVM, RF, NN, PCR, and BGLM). Before averaging across classifiers, the prediction scores were integrated across repetitions and CV folds. Each column corresponds to a certain signature type, which may be composed of different modules. The combined signature contains all predictive features from the mRNA, miRNA, and protein signatures. MI and PE indicate the additional use of molecular interaction and pathway enrichment features, respectively. Bar plots were generated for (A) C vs. NC classification, (B) NGC vs. GC classification, and (C) NGC vs. NC classification. Figure from Römer *et al.* (2014a).

across all three contrasts, which we did not observe after the combination of the multiple platform features and the integration of MI and PE features. For example, the mRNA and miRNA signatures performed much worse in the NGC vs. GC contrast than in the other two contrasts, whereas the protein signature failed to achieve a high AUC in the C vs. NC contrast. This indicates that the classification robustness can be increased by the integration of multiple molecular levels of gene regulation and an abstraction from molecular level to systemic, pathway level by calculating integrative features. Overall, we observed high (AUC > 0.95) and robust AUCs for the combined signature and the hybrid signatures complemented by the MI and PE features independent of the evaluated class contrast, whereas the single-platform features showed a strong dependence on the class contrast.

6.3.2 Predictive features for toxicogenomics models

To further analyze the performance and interpretability of the molecular signatures, we build consensus signatures by merging the feature rankings that we had obtained in each of the 10 repetitions. To this end, features were ranked by their average rank in the 10 repetitions. The optimal signature size was estimated as described above and for each classifier the consensus signature was created by selecting the top features from the feature ranking until the optimal size was reached. As an example, heat maps which illustrate the expression changes for the most informative genes or probes (for mRNA features), miRNAs, and proteins for distinguishing Cs from NCs are shown in Fig. 6.4. We observed a good correlation (Spearman's $\rho > 0.5$) between the change in gene expression and carcinogenicity classification for the mRNA features. For miRNAs and proteins, only a few molecules show a clear, carcinogen-specific expression. Particularly, PB and PBO show distinct expression patterns for some of the top-ranked miRNA and protein features (e.g., rno-miR-34a and CYP2C8, see Fig. 6.4B,C) This is consistent with the observed lower classification performance for miRNA and protein signatures in the C vs. NC contrast (see also Fig. 6.3).

We performed a literature search for the identified biomarkers to see if the signatures include genes, miRNAs, or proteins which have been linked to carcinogen exposure by other groups. In the mRNA signature for C vs. NC discrimination, we found many genes that are related to carcinogenic exposure, e.g., Gsta5 (Hayes *et al.*, 1998), Aldh1a1 (McMillian *et al.*, 2004), Ephx1 (Yates *et al.*, 2006), and Akr7a3 (Dewa *et al.*, 2009). These genes are mostly related to detoxification in the context of oxidative stress. Several probe sets that were not annotated with a rat gene are also included in the top features, but could not be verified against the literature. Two of the top miRNA markers for detection of carcinogen exposure, miR-34a (Dutta *et al.*, 2007) and rno-miR-200b (Tryndyak *et al.*, 2009), are associated with cancer formation. Several of the informative proteins for the C vs. NC contrast have also been linked to carcinogenesis, e.g., JUN (Sakai *et al.*, 1989), GLUL (DeBerardinis and Cheng, 2010), and CDKN1B (Nishimura *et al.*, 2008).

We also generated consensus signatures for the MI and PE features in the same man-

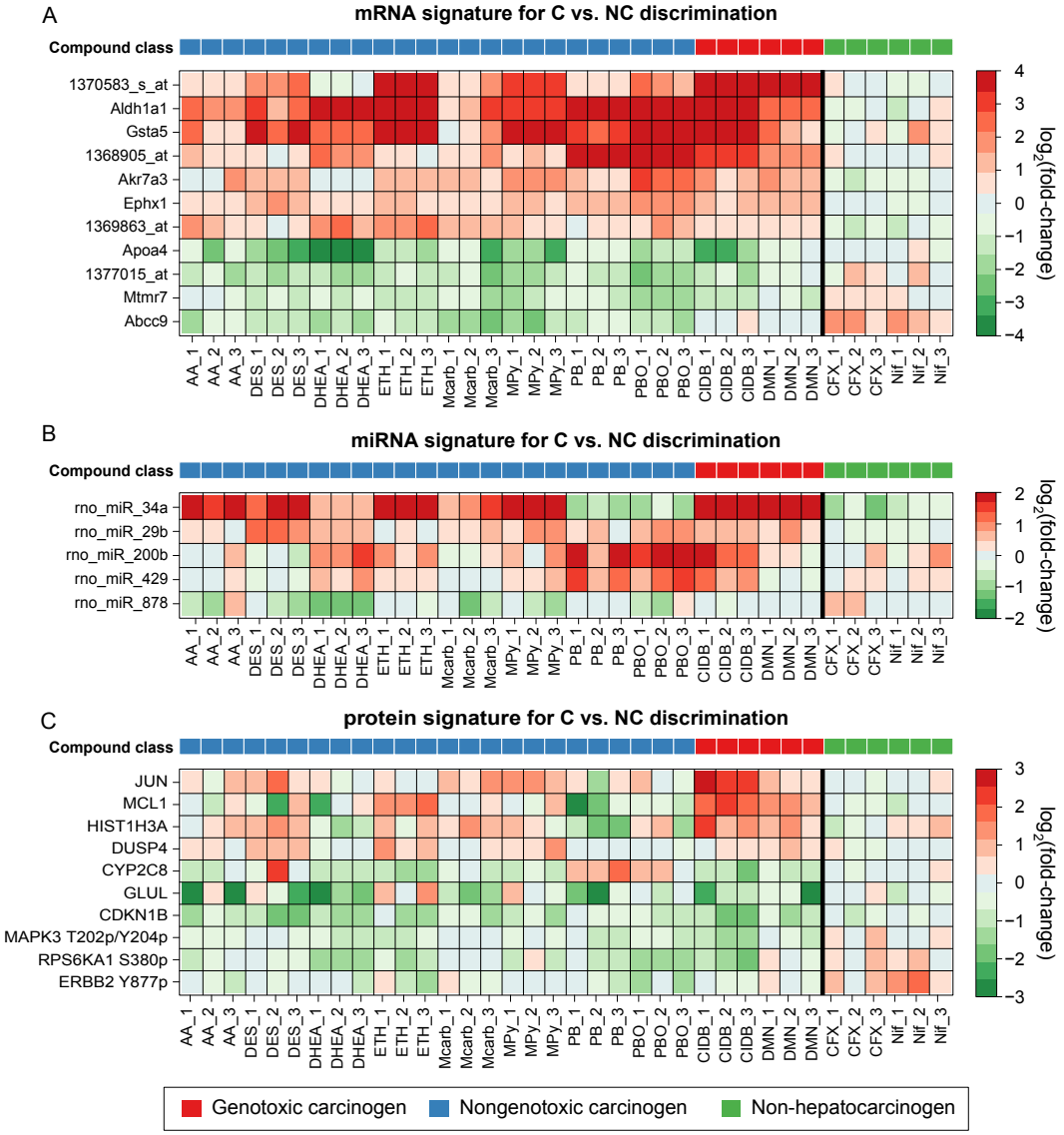


Figure 6.4: **Heat map plots of single-platform signatures for C vs. NC classification.** These heat maps depict characteristic expression patterns observed in livers of rats after exposure to several rodent liver carcinogens and noncarcinogens. A selection of signature molecules is shown for each profiled molecular level: (A) mRNA expression, (B) miRNA expression, and (C) protein expression. In each heat map, rows correspond to signature molecules and columns correspond to liver samples from differentially treated rats. The bold vertical lines separate the carcinogens from the noncarcinogens. Plotted are the $\log_2$(fold changes), where red indicates up-regulation and green indicates down-regulation (see color keys). The color bar on top refers to the compound class (see legend). Figure from Römer *et al.* (2014a).

ner as for the single-platform features. The most informative pathways for the C vs. NC and NGC vs. GC contrasts are shown in the heat maps in Fig. 6.5. The top pathways selected by SVM-RFE are highly specific for carcinogen exposure. Whereas no pathway is significantly enriched in all samples, none of the top pathways shows any enrichment in the NCs, which provides a clear, visual separation of the classes (Fig. 6.5A). In the NGC vs. GC contrast, the top features for discrimination of the classes also show very specific patterns, in particular for the genotoxic response to GC exposure (Fig. 6.5B). While few pathways are specifically enriched in NGCs, other pathways are enriched only in GCs, e.g., p53 related pathways. Only one of the three samples of the genotoxic substance DMN shows an uncharacteristic enrichment pattern. The enrichment of various pathways that are related to p53, which is a key gene in the cellular DNA damage response, is consistent with the expected DNA damage response after administration of genotoxic substances (Lopes *et al.*, 1997). Among the NGC-specific pathways are cytokine and interferon signaling pathways, which have been associated with nongenotoxic carcinogenesis before (Roberts and Kimber, 1999).

Figure 6.6 shows the interacting molecules of the MI feature signature for C vs. NC classification for selected compounds. In these volcano plots, we highlighted interactions where we observed a 1.5 fold up- or downregulation of both interacting partners, e.g., a miRNA and its mRNA target. Due to the biological interactions of the biomolecules, we expected that the expression of miRNAs and their target mRNAs are negatively correlated, whereas a positive correlation should be present for the expression of mRNAs and translated proteins. Several of the interactions in the MI signature involve genes that have been linked to carcinogenesis in the rat such as Glul (DeBerardinis and Cheng, 2010), Dusp1 (Feo *et al.*, 2009), Jun (Sakai *et al.*, 1989), Sgk1 (Won *et al.*, 2009), and Mgat4b (Liu *et al.*, 2010). Some of these genes have also been encompassed by the single-platform signatures that we discussed above. For the mRNAs of the genes Glul and Jun and their corresponding proteins, we found matching changes at transcriptional and translational levels after treatment with liver carcinogens (see Fig. 6.6). For NGCs, we observed a highly specific putative interaction between the miRNA rno-miR-29b and its potential target mRNAs Sgk1 and Mgat4. In contrast, we observed no putative interactions that were affected on multiple layers for NCs.

6.3.3 Toxicogenomics-based classification of undefined compounds

Of the 15 compounds that were selected for this exploratory study, the three compounds CPA, TAA, and WY are generally considered to be nongenotoxic carcinogens in the literature (for example in Ellinger-Ziegelbauer *et al.* (2008)). However, due to ongoing discussions on a possible genotoxic component to their carcinogenic potential, we decided to remove them from the training set to prevent confounding effects. For CPA, Lang and Redmann (1979) reported negative results in the Ames test, which indicates the absence of genotoxic activity, but later Martelli *et al.* (1996) reported positive results in a micronucleus test in female rats, such that the genotoxicity of CPA is still subject

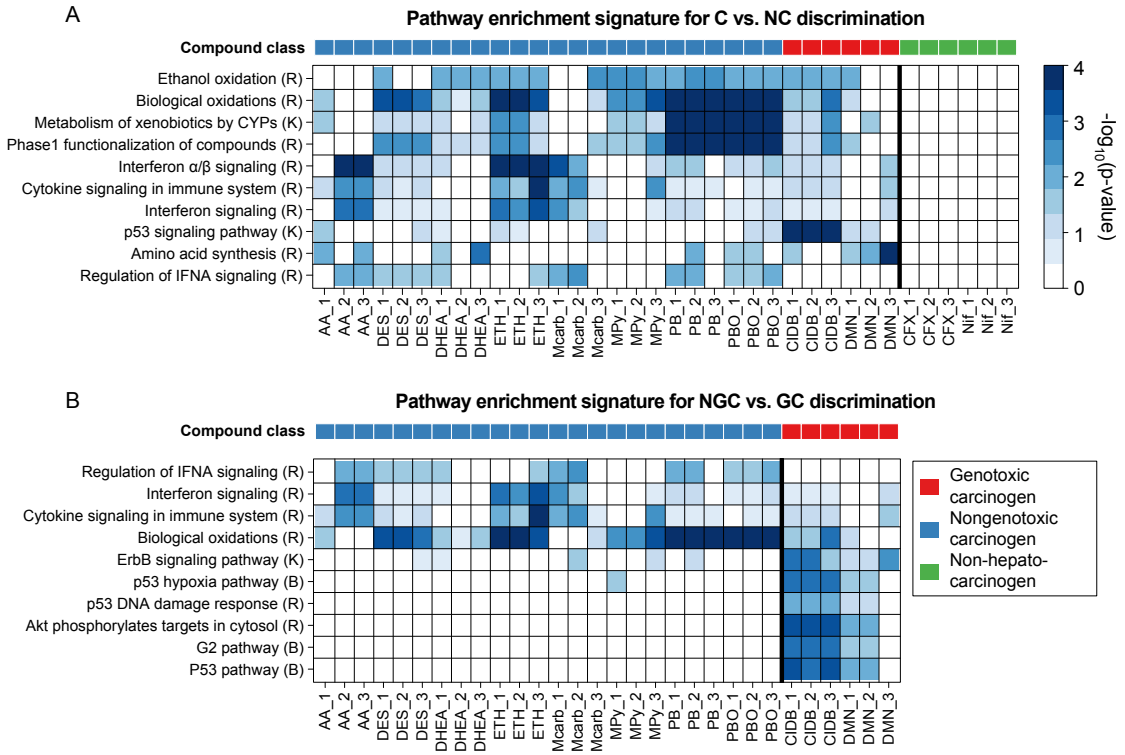


Figure 6.5: **Heat map plots of pathway enrichment signatures.** These heat maps depict overrepresentation of genes involved in relevant pathways among the genes deregulated in liver upon treatment of rats with a certain compound. Pathways relevant for compound classification were selected by SVM-RFE for different class contrasts: (A) C vs. NC and (B) NGC vs. GC. The rows correspond to canonical pathways from the databases Reactome (R), KEGG (K), or BioCarta (B) and the columns correspond to samples. The bold vertical lines separate carcinogens and NCs. The color of each cell refers to the $-\log_{10}(p\text{-value})$ obtained from a hypergeometric overrepresentation test and indicates the significance of a certain pathway enrichment (see color key). The color bar on top of each heat map denotes the carcinogenic class (see legend). Figure from Römer *et al.* (2014a).

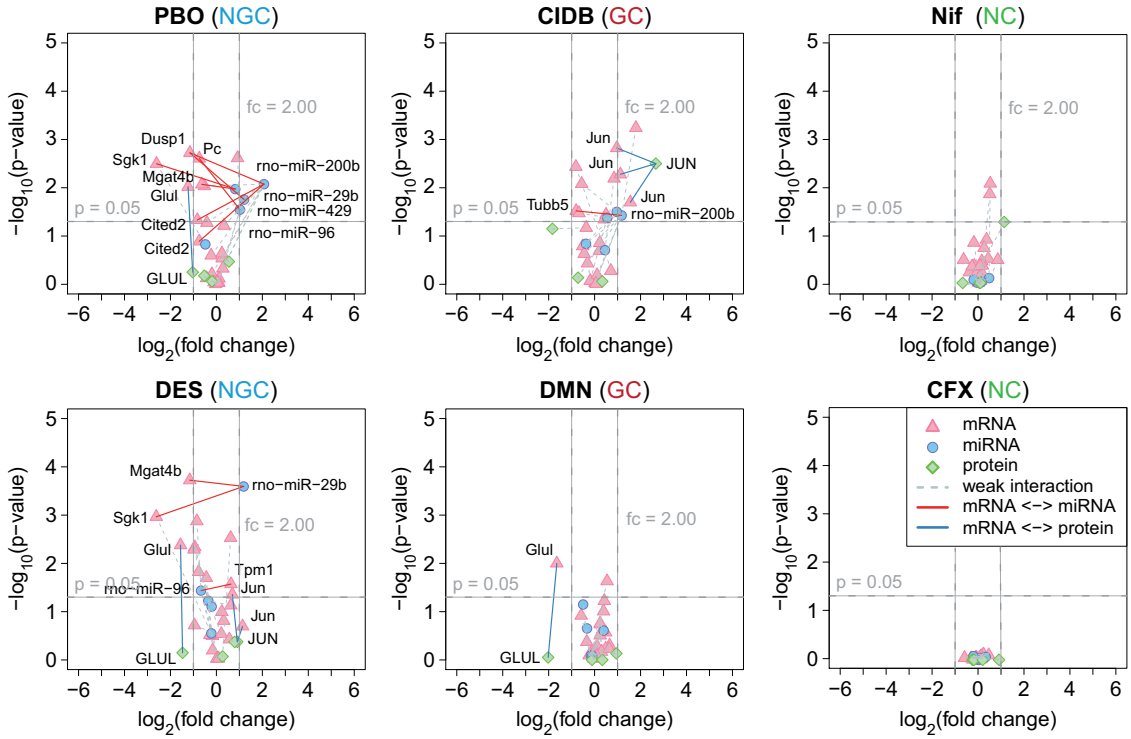


Figure 6.6: **Volcano plots of molecular interaction signatures.** Shown are volcano plots for two representative compound profiles of each of the three compound classes (i.e., NGC, GC, and NC). The plots represent putative molecular interactions between different molecular layers, which were found to be predictive for C vs. NC classification. For each interacting molecule (i.e., mRNA, miRNA, or protein) the strength of its differential expression was assessed in terms of the $\log_2$(fold change) and plotted against its significance, which is given by the FDR-corrected $\log_{10}(p\text{-value})$ obtained from a moderated t-test. Different shapes and colors denote different types of molecules (see legend). Colored edges were used to highlight molecular interactions for which a positive or negative correlation was observed between two molecule types. We considered correlations in the expression profiles of miRNAs and their experimentally confirmed or predicted mRNA targets, as well as between mRNAs and proteins sharing the same genomic locus. As a formal criterion for a putative molecular interaction, we required a 1.5-fold up- or downregulation for both interaction partners. Figure from Römer *et al.* (2014a).

of discussion. Similarly, we found negative Ames test results (Chieli *et al.*, 1987) and positive micronucleus tests (Mirkova, 1994) for TAA in the literature. WY is generally classified as a peroxisome proliferator, which is a subclass of nongenotoxic rodent hepatocarcinogens (Cattley and Popp, 1989), but Deutsch *et al.* (2001) reported positive results for a Comet assay with WY and Lefevre *et al.* (1994) observed clastogenicity in two cell types. These conflicting results led us to consider these compounds as not reliably labeled and we excluded them for all training and validation purposes.

To provide a mechanistic classification of these compounds and their putative mode of action for carcinogenicity, we used the signatures that were extracted from the remaining, well-annotated compounds for a predictive analysis with the same five classification methods that we used in the performance evaluation above. We also performed PCA to visualize the complex, high-dimensional expression patterns for all samples. The vector of signature features for each compound was transformed into a two-dimensional space spanned by the two principal components explaining most of the variance in the data and the resulting plots were generated for two different signatures for NGC vs. GC discrimination. The PCA plots for the mRNA signature and the hybrid signature, which encompasses all *omics* platforms and the MI and PE features, are shown in Fig. 6.7A and B. Most previous toxicogenomics studies used mRNA data and published mRNA signatures, which is why we compare the hybrid signature to the mRNA signature. Figure 6.7A shows a separation of the three classes based solely on mRNA expression changes. However, the two carcinogenic classes (GC and NGC) are rather close to each other and some samples treated with WY are placed outside of the space spanned by the NGC class. In contrast, the PCA plot for the hybrid signature with all *omics* signatures and the integrative MI and PE features in figure Fig. 6.7B shows a better separation of the three classes and all of the unclassified samples are encompassed by the area spanned by the NGC samples. This provides a clear indication for the classification of the undefined samples and compounds. Figure 6.7C-D shows a heat map of the classification confidence with the five classification methods that we used in the validation process. With the mRNA data, we observed that the unclassified samples are consistently classified as NGCs for most classification methods, indicating that the compounds CPA, TAA, and WY are indeed NGCs. When we used only the protein or miRNA signatures for the classification, we observed inconsistent classification and a low mean confidence for all samples. Using the hybrid signature with all available feature signatures, we observed high confidence scores and a robust classification across all employed learning algorithms, which further supports our claim that the integration of multiple *omics* features and complex, integrative features enhances the classification performance. All undefined compounds were classified as NGCs based on their molecular profiles. Our results thus support the outcomes of the carcinogenicity assays that were performed for WY, TAA, and CPA (Lang and Redmann, 1979; Chieli *et al.*, 1987; Cattley and Popp, 1989).

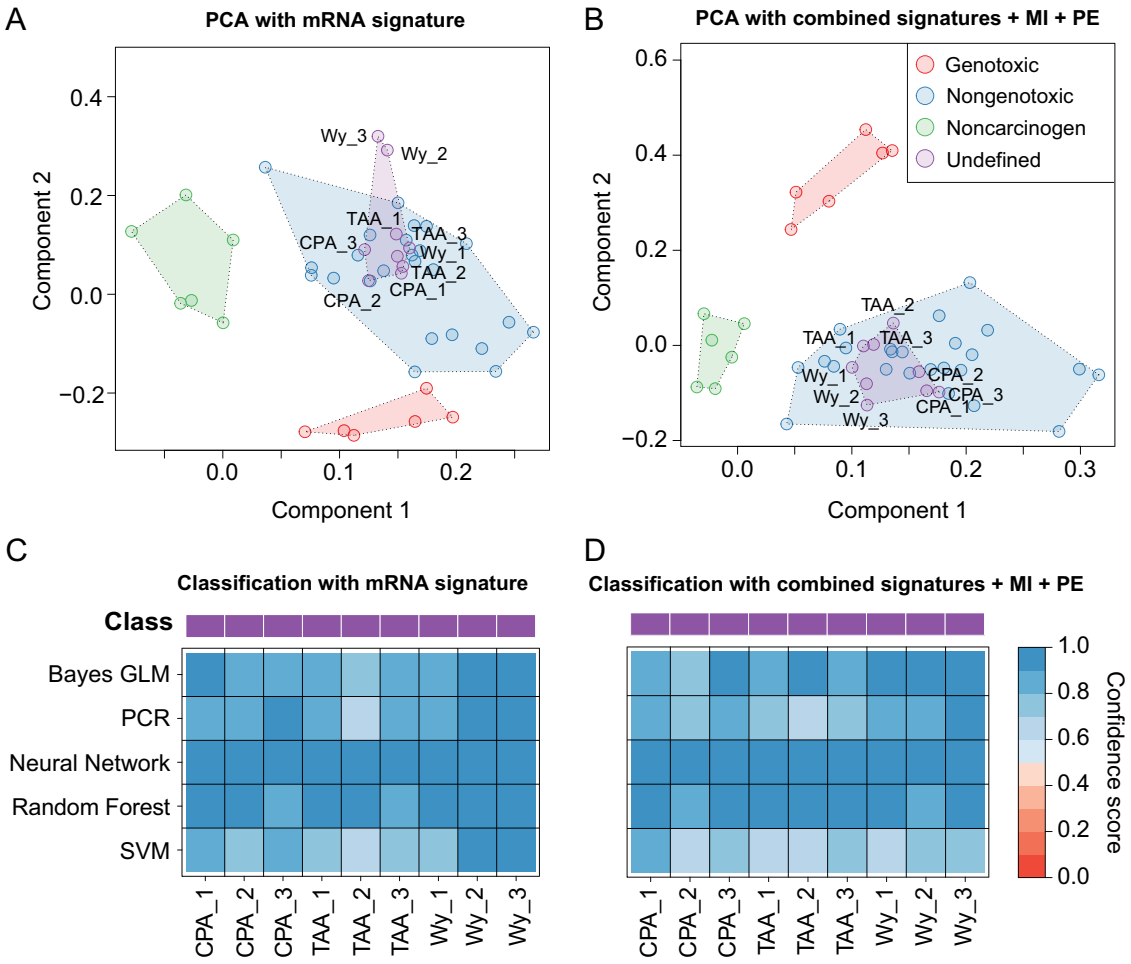


Figure 6.7: **Classification of undefined compounds.** (A) Samples are represented based on the mRNA signature for NGC vs. GC discrimination. The corresponding fold changes were PCA-transformed and plotted in a lower-dimensional space spanned by the first two principal components. The color of the spheres corresponds to the class of the administered compounds with respect to their hepatocarcinogenic properties in rats. Clusters of rat liver samples after treatment of rats with compounds of the same class are highlighted by means of transparent polygons. (B) Same as (A), but instead of using the mRNA signature for sample representation, all single-platform (mRNA, miRNA, protein) and cross-platform signatures (PE, MI) were combined. (C) The heat map displays the confidence scores obtained from five different machine learning methods that were used to classify the undefined compounds CPA, TAA, and WY as either NGC or GC. The confidence scores are [0, 1]-scaled and correspond to the probability that a certain sample was treated with an NGC (see color key). (D) Similar illustration as in (C) obtained with classifiers trained on all signatures combined. Figure from Römer *et al.* (2014a).

6.4 Summary and conclusions

In this chapter, we proposed a method for the prediction of carcinogenic effects of chemicals in the liver of rats in the LRB. The proposed method is based on expression data, which was obtained by profiling changes in the expression of mRNAs, miRNAs, and proteins using multiple *omics* technologies after rats received a daily oral administration of a set of prototypic Cs and NCs for up to 14 days. The development of short-term (two to four weeks) assays that reliably detect putative carcinogens can facilitate and accelerate the development of drugs and industrial chemicals. New short-term assays could be used to perform a prescreening of candidate chemicals to prioritize those chemicals that are less likely to induce tumors for LRBs or even replace the LRB. Toxicogenomics is currently one of the most promising approaches for the development of such short-term assays. Most of the toxicogenomics studies that have been performed previously profiled the mRNA expression changes and employed machine learning to infer signatures and build predictive models (Ellinger-Ziegelbauer *et al.*, 2008; Uehara *et al.*, 2011; Auerbach *et al.*, 2010). Some groups have also tried to use other, more recent *omics* technologies such as miRNA or protein expression changes to build predictive models (Schmitz-Spanke and Rettenmeier, 2011; Yokoi and Nakajima, 2011). Here, we used not only the data from one *omics* layer but included and integrated multiple layers of *omics* data to increase the predictive power of the prediction models. In addition to the profiled expression changes, we further introduced two integrative feature types (molecular interactions and the pathway enrichment), which use prior knowledge from interaction and pathway databases to calculate new features for model building.

We used a repeated nested-CV workflow to evaluate the predictive power of signatures that were inferred using either only the single *omics* platforms, the combination of all *omics* platforms, or the integrative features. For this evaluation, a unique dataset was generated, in which mRNA microarrays, miRNA microarrays, and RPPAs were used to profile expression changes on three *omics* levels for 15 prototypic compounds. The dataset was deposited in GEO and is available under the accession number GSE53085 (Edgar *et al.*, 2002; Barrett *et al.*, 2013). We used 12 well-annotated compounds to construct models with an unbiased 2×2 nested CV with 10 randomly created splits into training and validation set. The three remaining compounds had conflicting genotoxicity information and were therefore excluded from the training and validation sets.

We observed that the predictive power of classification models was higher when we used the combination of all three *omics* levels compared to models built based on a single *omics* platform. This observation was consistent across the three different class contrasts for which we built models (C vs. NC, NGC vs. GC, and NGC vs. NC). The predictive power could be increased further by the inclusion of the integrative MI and PE features. These features provide an abstraction of the molecular expression changes to molecular interactions and biochemical pathways. We observed a consistent increase in the average AUC that was obtained for each signature with five different classification methods, with the exception of the NGC vs. GC contrast, where the protein signature

performed slightly better. The inclusion of the integrative features also improved the robustness by reducing the variance in classification AUC in the random resamplings. We also observed a better separation of the three classes in a PCA plot when using the hybrid signature compared to single-platform signatures. In summary, the combination of multiple *omics* layers yielded a high prediction accuracy (AUCs > 0.9) in all class contrasts and outperformed the single-platform signatures in terms of power and robustness. The best classification (AUCs > 0.95) was achieved by including the pathway enrichment and molecular interaction features.

We used the single-platform, combined multi-platform, and hybrid signatures to reclassify the three compounds CPA, TAA, and WY for which conflicting genotoxicity information was found in the literature. These three compounds are known to be carcinogenic in rats and have been negative in Ames tests, but have shown some positive results in micronucleus or Comet assays (e.g., Lang and Redmann, 1979; Martelli *et al.*, 1996; Chieli *et al.*, 1987; Mirkova, 1994). Our prediction models have consistently predicted all three substances to be nongenotoxic carcinogens, which is consistent with their classification by other experts (e.g., Ellinger-Ziegelbauer *et al.*, 2008; Uehara *et al.*, 2011). In line with the results of the model performance evaluation, we observed a much higher prediction confidence for the hybrid signatures compared with the single-platform signatures.

In conclusion, this exploratory study demonstrated that the combination and integration of data from multiple layers of gene regulation improves the power of prediction models in toxicogenomics. Studies with more compounds and global profiling technologies, e.g., for protein expression and DNA methylation, are necessary to fully assess the impact of multi-*omics* data in the context of toxicological risk assessment. We believe that future toxicogenomics studies can benefit from profiling additional *omics* layers, e.g., metabolomics, DNA mutations, or genome-wide promoter methylation. We hope that our work encourages the maintainers of the currently available, large databases in toxicogenomics (e.g., TG-GATEs and DrugMatrix) to generate additional data to complement the existing mRNA data.

Chapter 7

Web-based interactive visualization of gene regulation data

The detection of tumors during preclinical toxicology studies in drug development can result in the delay of drug candidates reaching the market depending on the MOA, human relevance, and the intended therapeutic indication. Drug candidates associated with direct genotoxicity leading to DNA mutations are efficiently eliminated early during development with established short-term assays. However, NGCs form a significant proportion of carcinogenic drug candidates and induce tumors through mechanisms other than DNA mutations. To identify such compounds, time-consuming *in vivo* LRBs are required. These LRBs require a large number of animals, take at least three years to be completed, and cost up to 2 million US dollars (Johnson, 2012). The earlier detection of cancer development upon compound treatment could lead to significant savings in time, cost, and animal numbers, and could benefit patients regarding drug safety.

In this context, the MARCAR project (MARCAR Consortium, 2010) generated a wealth of data to increase insights into mechanisms of NGC action to detect carcinogenic effects of drug candidates earlier (e.g., Lempiainen *et al.*, 2013; Unterberger *et al.*, 2014; Thomson *et al.*, 2014; Römer *et al.*, 2014a). The MARCAR project explored primarily the liver, which is the major target organ of NGC induced tumors in rodents (Knight *et al.*, 2006a). A particular goal was the identification of early biomarkers of NGC effects to provide the basis for designing new short-term assays for earlier detection of potential NGCs. GCs and NCs were used as control substances to ensure that identified biomarkers are specific for NGCs.

A fundamental concept of the MARCAR project was the integration of data collected at multiple levels of gene expression regulation. In addition to traditional profiling of mRNA abundance with microarrays, non-coding mRNAs were measured by array technology, protein abundance and modifications were assessed by RPPAs, and DNA methylation patterns were identified with methylation arrays (Unterberger *et al.*, 2014; Römer *et al.*, 2014a; Thomson *et al.*, 2012). Potential biomarkers identified include, e.g., the *Dlk1-Dio3* imprinted gene cluster of noncoding RNAs (Lempiainen *et al.*, 2013), epigenetic changes in the 5-hydroxymethylome (Thomson *et al.*, 2012), and effects on the hepatic mesenchyme (Riegler *et al.*, 2015). To effectively explore the interaction of mul-

tiple layers of gene regulation, the InCroMAP software was developed, which provides KEGG pathway maps with overlaid information on, e.g., mRNA expression, miRNA-gene interactions, and methylation changes (Wrzodek *et al.*, 2013). Toxicogenomics approaches were applied to extract signatures (i.e., lists of biomarkers) to identify NGC action in both rats and mice (Kossler *et al.*, 2015; Eichner *et al.*, 2013b). In total, the MARCAR project has generated 27 datasets across three species (*Homo sapiens*, *Rattus norvegicus*, *Mus musculus*) and four regulation levels (mRNAs, miRNAs, proteins, and DNA methylation), which are available from GEO under accession number GSE68387.

The majority of MARCAR datasets were generated with traditional mRNA microarray profiling techniques. These arrays measure the expression of almost all known genes in the genome of a certain species at once. This provides an enormous amount of raw data that requires preprocessing, normalization, and statistical or computational approaches to effectively identify patterns and specific biomarkers. Traditionally, bioinformaticians would process the data and provide tables and static visualizations to biologists, who would then try to interpret the data. To facilitate the process of generating insights from the wealth of data, we created MARCARviz, a set of software tools that allows an interactive visual analysis to quickly identify relevant patterns in large amounts of microarray data by non-bioinformaticians. MARCARviz can be used to quickly answer common questions associated with the experiments, e.g., which genes are affected by the treatment with a specific compound, generate a mechanistic hypothesis, for example by finding enriched pathways and Gene Ontology (GO) terms, or compare the effects of several compounds. MARCARviz also provides cross-platform and cross-species analysis, which usually require the mapping of several types of identifiers (e.g., Affymetrix probe IDs to gene names) or identification of orthologous genes. The most relevant visualizations for microarray data (e.g., heat maps, Venn diagrams, volcano plots) are provided and extended with interactive functionality for pattern identification.

MARCARviz is available as a web platform, which requires no additional plugins and few computational resources on the user's side. All analyses are performed on a computation cluster to provide results as fast as possible. The major advantage of MARCARviz is that it allows biologists without advanced bioinformatics knowledge to quickly answer the most common questions that might be asked of the MARCAR mRNA expression data, extract data and figures that support their hypothesis, and generate new insights and hypotheses about NGC mechanisms and biomarkers. This chapter was presented at the German Conference for Bioinformatics 2016 and is available as a preprint from *PeerJ Preprints* (Römer *et al.*, 2016a).

7.1 Database content and construction

7.1.1 Datasets and data analysis

All animal experiments have been approved by the respective ethics committees and were performed according to established experimental guidelines. Study design and raw data generation have been described in previous publications (Unterberger *et al.*, 2014; Riegler *et al.*, 2015; Braeuning *et al.*, 2010; Luisier *et al.*, 2014; Eichner *et al.*, 2014a; Lempiäinen *et al.*, 2011; Ellinger-Ziegelbauer *et al.*, 2008; Braeuning *et al.*, 2016). The datasets cover three species: mouse (*Mus musculus*), rat (*Rattus norvegicus*), and humans (*Homo sapiens*). The MARCAR studies can be grouped into two broad categories: (i) short-term effects of NGCs and (ii) mechanistic analysis of a model NGC (phenobarbital). To study the common characteristics of NGCs, rats and mice were administered daily doses of several NGCs for up to four weeks. For comparison, GCs and NCs have been included to identify effects that are specific for NGCs (see Table A.3). For the mechanistic investigation of the model substance phenobarbital, several knockout studies have been performed in mice, along with transcriptomic profiling of tumor and normal tissue. In total, 16 microarray datasets have been generated and processed for inclusion in MARCARviz. All raw data was submitted to GEO and is available under the accession number GSE68387 (see Table A.4).

Microarray quality control was done with the R package `arrayQualityMetrics` to remove outliers and low-quality samples (Kauffmann *et al.*, 2009). Raw data processing was performed with the R packages `affy` and `limma` (Gautier *et al.*, 2004; Smyth, 2005). In short, Affymetrix 3' IVT expression array data was normalized with RMA and summarized to Entrez Gene IDs using custom Brainarray CDF files (Dai *et al.*, 2005). Agilent microarray data was within-array background corrected, quantile normalized between arrays, and summarized to Entrez Gene IDs. For each dataset, we normalized all samples together. To eliminate batch effects, all studies use designs that ensure that treated and control samples are run in a single batch. A moderated t-value implemented in `limma` was used to compute p-values for significant deregulation of genes. We used the Benjamini-Hochberg method to correct for multiple hypothesis testing. Logarithmized (base 2) fold changes were calculated as the $\log_2$ of the observed mean intensity ratio between treated and control animals for each gene. The `biomaRt` package for R was used to identify orthologous genes for cross-species data comparison (Durinck *et al.*, 2009).

Gene to gene set mapping files were obtained from the Molecular Signature Database (MSigDB, Liberzon *et al.* (2011)) for the KEGG, BioCarta, Reactome, and Gene Ontology databases (Ashburner *et al.*, 2000; Nishimura, 2001; Matthews *et al.*, 2009; Kanehisa *et al.*, 2012). For the enrichment analysis, all gene identifiers are mapped to orthologous HGNC gene symbols based on information available from the Entrez Gene database (Maglott, 2004). A hypergeometric test was used to calculate enrichment p-values. All calculated p-values were corrected for multiple hypothesis testing with the method developed by Benjamini-Hochberg.

7.1.2 Database construction

We used a combination of spreadsheets, R scripts, and a schema-free NoSQL database to edit, process, and store all data and metadata. The biggest metadata unit is called `Study`, which is conceptually similar to a GEO entry. A `Study` usually encompasses several different conditions (called `Treatment`) that are compared to identify biological effects.

A `Treatment` constitutes a specific set of conditions that is of interest. For example, toxicogenomics studies often involve multiple drugs that are administered to animals of the same species for a specific time period. In this example, each drug constitutes a `Treatment`. If multiple parameters are varied in a study, e.g., multiple drugs are measured at several different time points, each `Treatment` corresponds to a unique combination of parameters. All MARCAR studies used a case-control design, i.e., for each `Treatment` a corresponding control is available, which establishes the baseline for comparison. Most MARCAR studies also used biological replicates to allow a statistical analysis of the observed biological effects. Thus, each `Treatment` encompasses several units called `Samples`, each of which is labeled as either treated or control `Sample`.

A `Sample` is the smallest unit of metadata and represents a single microarray experiment (e.g., a CEL file for Affymetrix microarrays). For each `Study`, this metadata scheme was represented by an XLS file with three tables: one table for the `Study` information, one table for the metadata of each `Treatment`, and one table for the `Sample` metadata that also contains the raw data file paths. We decided to use XLS spreadsheets because they facilitate the editing of the metadata, allow the storage of all three metadata levels in one file, and can easily be parsed with R.

We processed the raw microarray data as described above with a standardized R script. The R script read the metadata spreadsheet to determine the microarray platform, the experimental design, and the assignment of raw data files (i.e., `Samples`) to each `Treatment` for statistical analysis. The processed data was stored as an R workspace in binary format for faster reading. Also, the metadata tables were included in the R workspace to provide a single file that contains all relevant data and metadata. An R script pushed all metadata and data into the NoSQL database.

To store the processed microarray data, we used one table for each `Study` that was named `expr_[Study]`. With the NoSQL-based MongoDB database system[1], this allowed for an easy retrieval of expression data for individual genes or `Treatments` as well as for the whole expression table. The expression table was exported from R as a CSV file that is subsequently imported into the MongoDB. Each expression data table encompassed all three types of stored data: the normalized signal intensities for each probe, the fold change, and the p-value for differential expression. We created indices automatically to ensure a fast retrieval of expression data for the expected query structures.

[1] `https://www.mongodb.com/`, accessed 15 September, 2016

Seperate tables were created to store the metadata data for `Studies` and `Treatments`. Because sample metadata was only required for the data processing, the `Samples` are only included in the `Treatment` table. Due to the use of a schema-free database system, we could adapt the metadata tables to allow for additional fields in the XLS files. Through these additional fields, other users of the MARCARviz framework can include additional metadata fields without having to modify the database schema. For example, the `Study` table includes a field `details`, which includes all fields from the study table in the XLS file that are not required. This `details` field can be used to provide users with information on particular study designs, e.g., on the summary web page for a `Study`. We also created several database tables that were necessary for managing jobs and users. Although we used the schema-free MongoDB database system, we applied concepts of relational databases where applicable.

Figure 7.1 shows an approximate database schema of the final database.

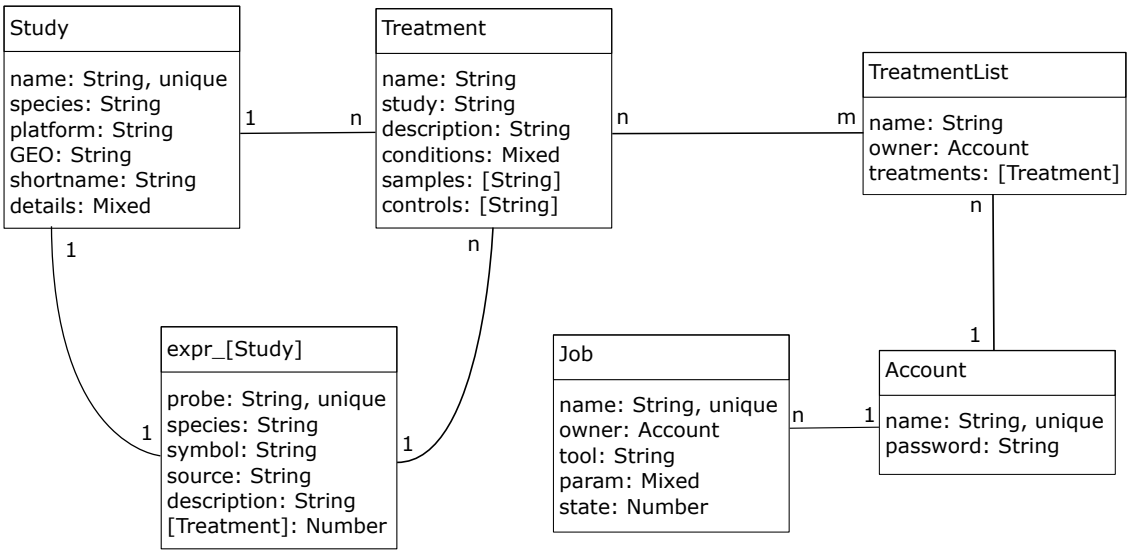


Figure 7.1: **Schema of the MARCARviz database.** Each MARCAR dataset is represented as a `Study`, which is identified by a unique internal identifier (`name`). A `Study` consists of several `Treatments`, each of which corresponds to a group of biological replicates (`samples` and `controls`). The fields `details` (`Study`) and `conditions` (`Treatment`) enable the storage of additional metadata in a schema-free manner. Expression data is stored in an individual table for each study to allow fast retrieval through MongoDB's document storage and querying concept. Indeces were created to allow a fast retrieval of data with common queries.

7.1.3 Architecture of the interactive web platform

The MARCARviz front end is written entirely in the current web standards HTML5, JavaScript, and CSS. In consequence, MARCARviz requires no Flash or Java applets to achieve its interactive functionality. We use several existing JavaScript libraries to generate the interactive visualizations directly in the browser. For scatter, volcano, bar, and box plots, we use Highcharts JS[2], a JavaScript library that provides interactive, highly-customizable charts. We use Highcharts box plots to visualize the distribution of expression values and bar plots to show the signal intensity across multiple replicates. Both volcano and scatter plots are a customized variant of Highcharts' *x*-*y*-plots. To create interactive heat maps, we used a modified version of InCHlib.js (Škuta *et al.*, 2014), an open-source JavaScript library for interactive heat maps based on KineticJS. We extended InCHlib.js with legends, which provide additional metadata for samples and are visible when exporting static images of the interactive heat map, e.g., for presentations or publications. Venn diagrams were created with jvenn.js (Bardou *et al.*, 2014), which is a plug-in for the popular jQuery JavaScript library. Interactive and responsive tables were created with the DataTables[3] plug-in for jQuery. The DataTables library extends standard HTML tables with functionality such as sorting, pagination, and searching. We used the Bootstrap[4] framework to customize the style of our web platform through CSS and JavaScript. Bootstrap is an open-source framework that is well integrated with many other libraries that we used.

As middleware, MARCARviz uses a Node.js[5] web server to handle communication between server and client. Node.js is a runtime environment for server-side web applications, which uses JavaScript for development. In particular, Node.js uses an asynchronous approach for processing input and outputs in an event-driven architecture to optimize performance and scalability. Thus, the Node.js server can handle simultaneous requests from multiple clients to provide real-time data responses to provide the data for the interactive visualizations. We used the Express framework[6] to implement the server-side application of our web platform. Express provides utility methods for HTTP request handling, e.g., parsing and session management, along with an API for other middlewares that facilitate client-server communication. All client-server traffic is encrypted according to the HTTPs standard with Express.

The Node.js server distributs requests to a backend that handles data retrieval from the MongoDB database and processes the data for the visualizations. The R backend preprocesses the expression data and exports JavaScript Object Notation (JSON) files, which are passed back to the client's browser and can easily be parsed and visualized with JavaScript. A schematic overview of the architecture is shown in Fig. 7.2.

[2] `http://www.highcharts.com/`, accessed 21 June, 2016
[3] `https://datatables.net/`, accessed 21 June, 2016
[4] `http://getbootstrap.com/`, accessed 21 June, 2016
[5] `https://nodejs.org/`, accessed 21 June, 2016
[6] `http://expressjs.com/`, accessed 21 June, 2016

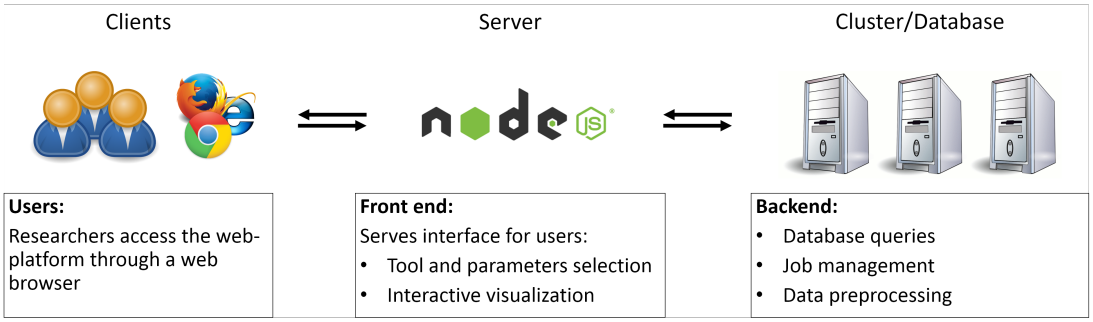


Figure 7.2: **Architecture of the MARCARviz web platform.** This is a simplified scheme of the MARCARviz platform architecture that demonstrates the interaction of the general components. Researchers access MARCARviz through the internet with the supported major browsers and use the web interface to request analyses or visualizations. A node.js server handles the communication, serves static content, and manages the distribution of the requested analyses to the computation cluster. Database queries, data preprocessing and integration (e.g., for data from multiple studies) are performed on the computation cluster. The results of the requested analyses are then returned to the user and rendered by the browser with Javascript, HTML5, and CSS.

7.2 Interactive visualizations of microarry data

The MARCARviz web platform is navigated through a menu available on all pages. We made a major distinction between "Analysis" and "Visualization" tools. Typically, an analysis will give a report in the form of a table, e.g., of differentially expressed genes or enriched pathways. A visualization will produce an interactive plot that can be used for visual exploration of the data or as a figure supporting a hypothesis in a manuscript or presentation. Currently, MARCARviz supports two analyses and five visualizations, which are described in the following sections along with possible use cases. An example of the user interface is shown in Fig. 7.3.

Differential expression tables

The differential expression analysis allows the identification of genes that are affected by a specific condition, e.g., after treatment of rodents with an NGC. Differential expression is established by comparing the gene expression in treated animals or tumor tissue with expression in untreated control animals or normal, non-tumor tissue. The strength of differential expression is measured as a fold change, i.e., the ratio of expression between treated and control animals. The significance of deregulation is given as the p-value of a moderated t-test. Users can define custom fold change and p-value thresholds to identify differentially expressed genes. The deregulated genes are shown in a table that also gives a summary for each gene and provides links to external databases with additional information on the gene. The user can also inspect the observed expression in individual

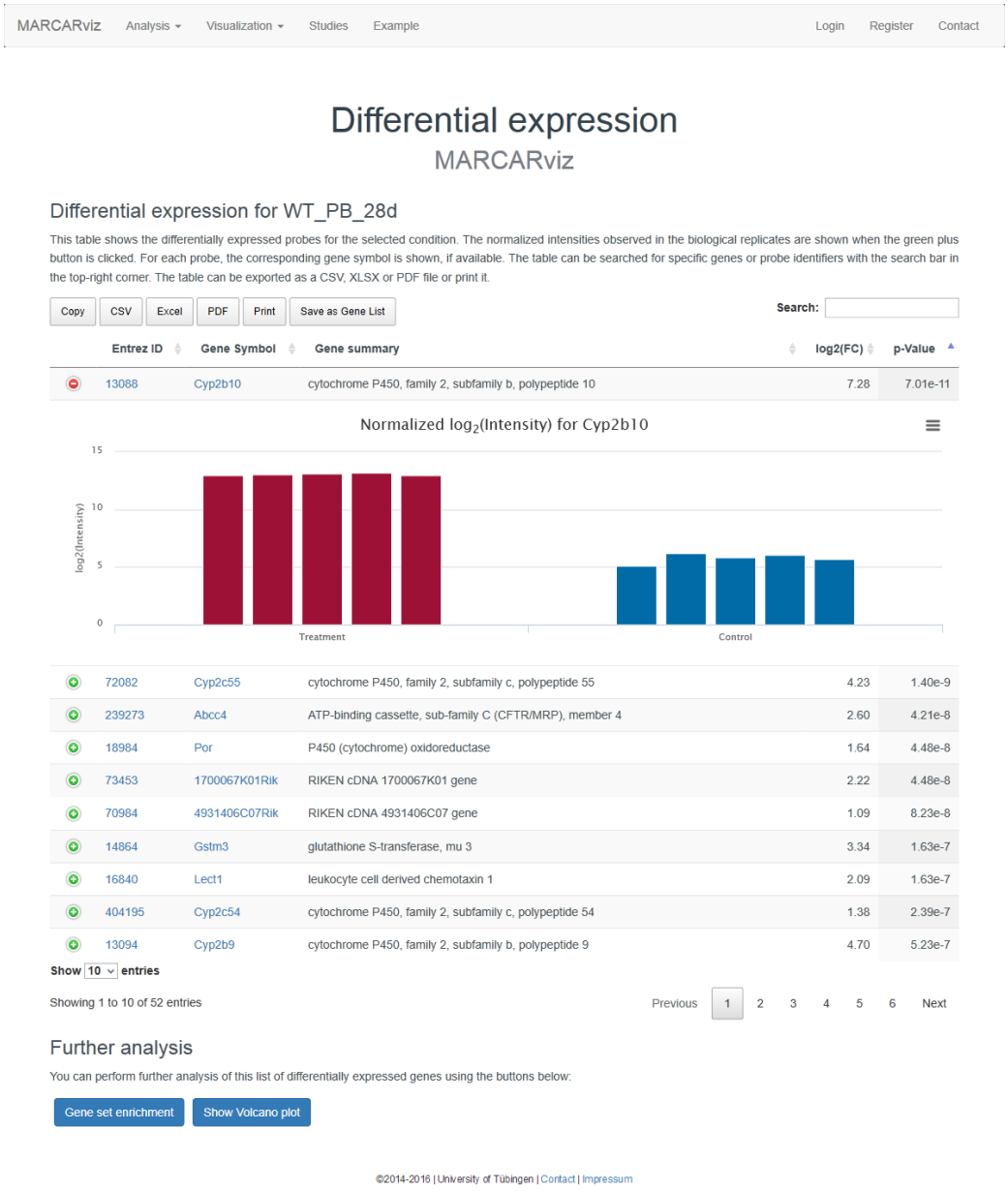


Figure 7.3: **User interface of an interactive table of differential expression.** This screenshot shows the result page of the differential expression tool. The differentially regulated genes that pass the inclusion criteria defined by the user are shown in a table. The table provides supplementary information, e.g., the official gene symbol and a short gene summary. By clicking on the table row, the expression data for the selected gene is shown in more detail. The table can be exported in common file formats (CSV, XLS, and PDF) or saved as a gene list for further analysis. Links to a gene set enrichment analysis and a volcano plot visualization of the data are also provided.

samples to confirm the observed deregulation. The list of genes identified by this tool can be saved to be used as input for other tools, e.g., for a gene set enrichment analysis, or exported as a CSV, PDF, or XLS file. Figure 7.4(a) shows a differential expression table created with MARCARviz.

Gene set enrichment analysis

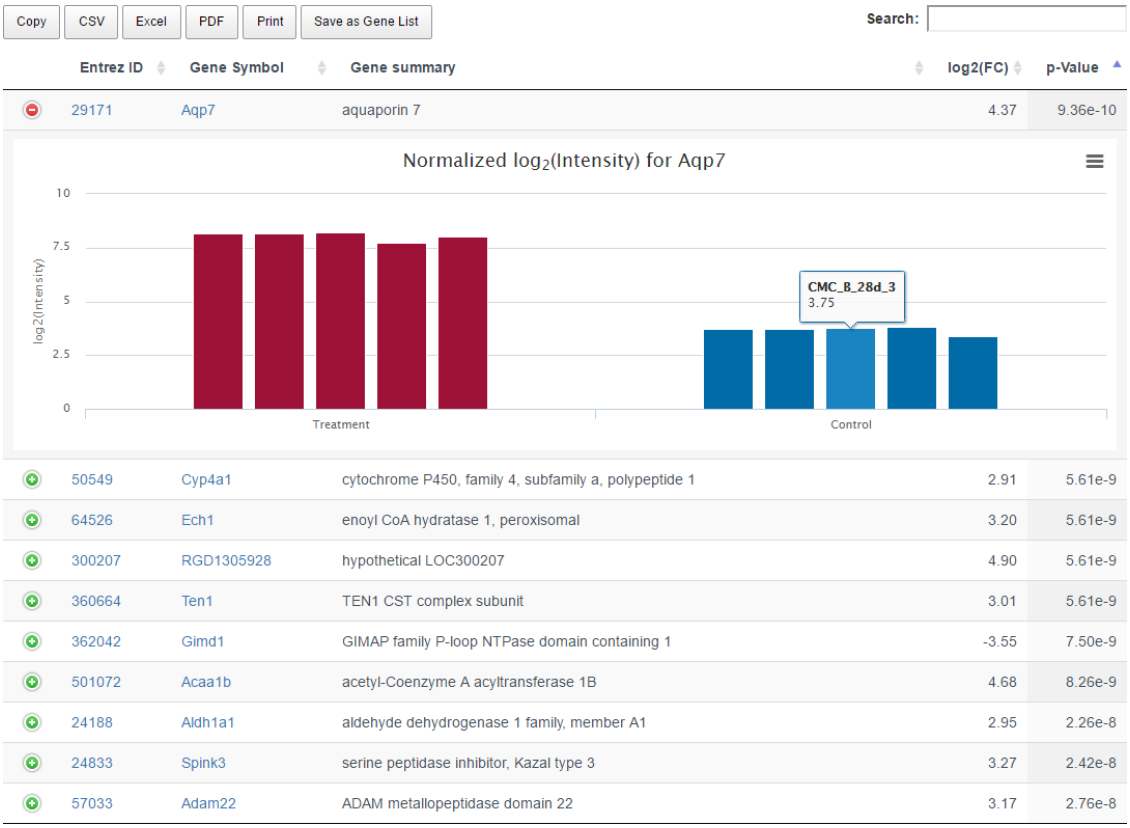
Gene set enrichment analyses identify gene sets, e.g., biochemical pathways or GO categories, for which a higher-than-expected number of genes is up- or downregulated in a condition. The test for enrichment is performed with a hypergeometric test and and the result is reported as an enrichment q-value, i.e., the Benjamini-Hochberg multiple-testing corrected p-value. The user can define the thresholds applied to filter differentially expressed genes and select the gene sets for which enrichment is tested. Currently, MARCARviz supports enrichment tests for GO categories and three pathway databases: KEGG, BioCarta, and Reactome. The results of the gene set enrichment analysis are shown in a table that reports the significance for each gene set along with the deregulated genes, the gene set statistics, and links to external databases with additional information on the gene sets. The table can be exported as a CSV, PDF, or XLS file

Volcano plots

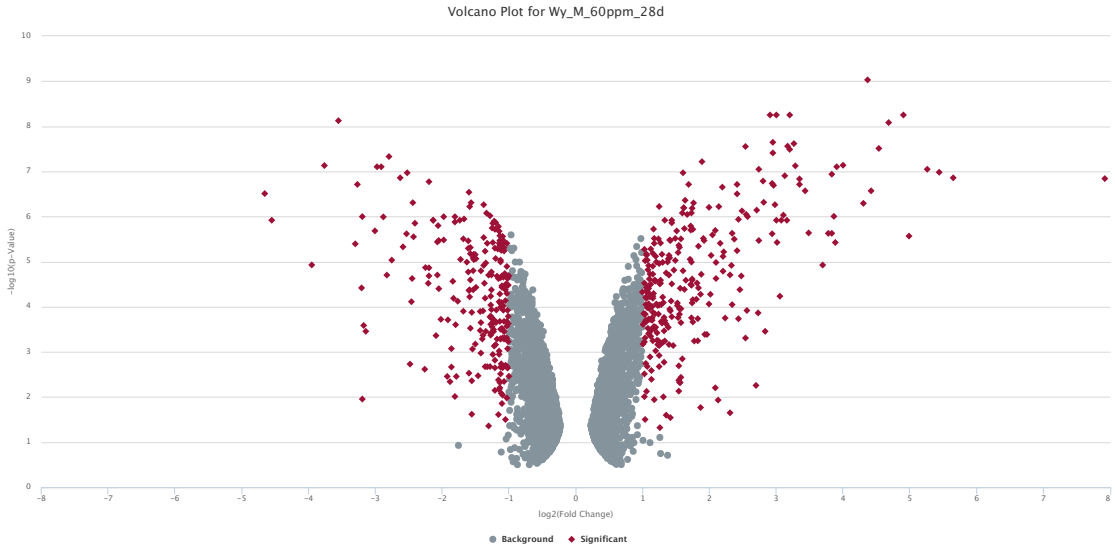
Volcano plots are a visual representation of the strength and significance of differential gene expression in a single condition. Fold change and p-value are used to represent the condition-dependent gene regulation. The interactive volcano plot shows the effect strength, i.e., the fold change, on the x-axis and the significance, i.e., the p-value, on the y-axis. Each point in the plot corresponds to a single gene. The gene symbol and the observed expression in samples and controls can be displayed by hovering and clicking on a point. Again, users can set individual thresholds for fold change and p-value and save the list of deregulated genes for further analyses. The volcano plot can be exported as PNG or PDF. Figure 7.4(b) shows a volcano plot created with MARCARviz.

Scatter plots

Scatter plots show the general effects on gene expression in two conditions. For each gene, the fold change in the first condition is plotted on the x-axis and the fold change in the second condition on the y-axis. In addition, a linear regression analysis is performed to provide statistical measures (for example the R^2 value) of the concordance between the conditions. This tool supports cross-platform and cross-species comparisons, e.g., to compare the effects of the administration of a carcinogenic substance in rats and mice. Users can set a fold change threshold to exclude non-affected genes from the comparison and export the resulting plot as PNG or PDF.

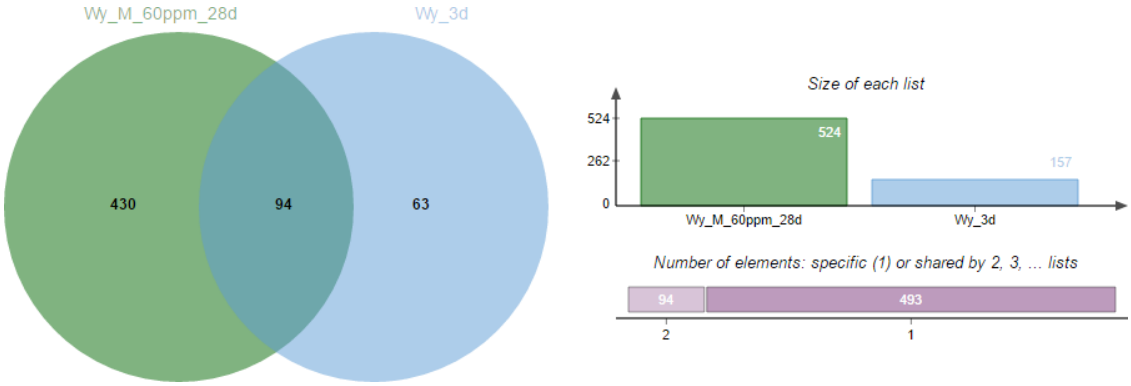


(a) Differential expression table

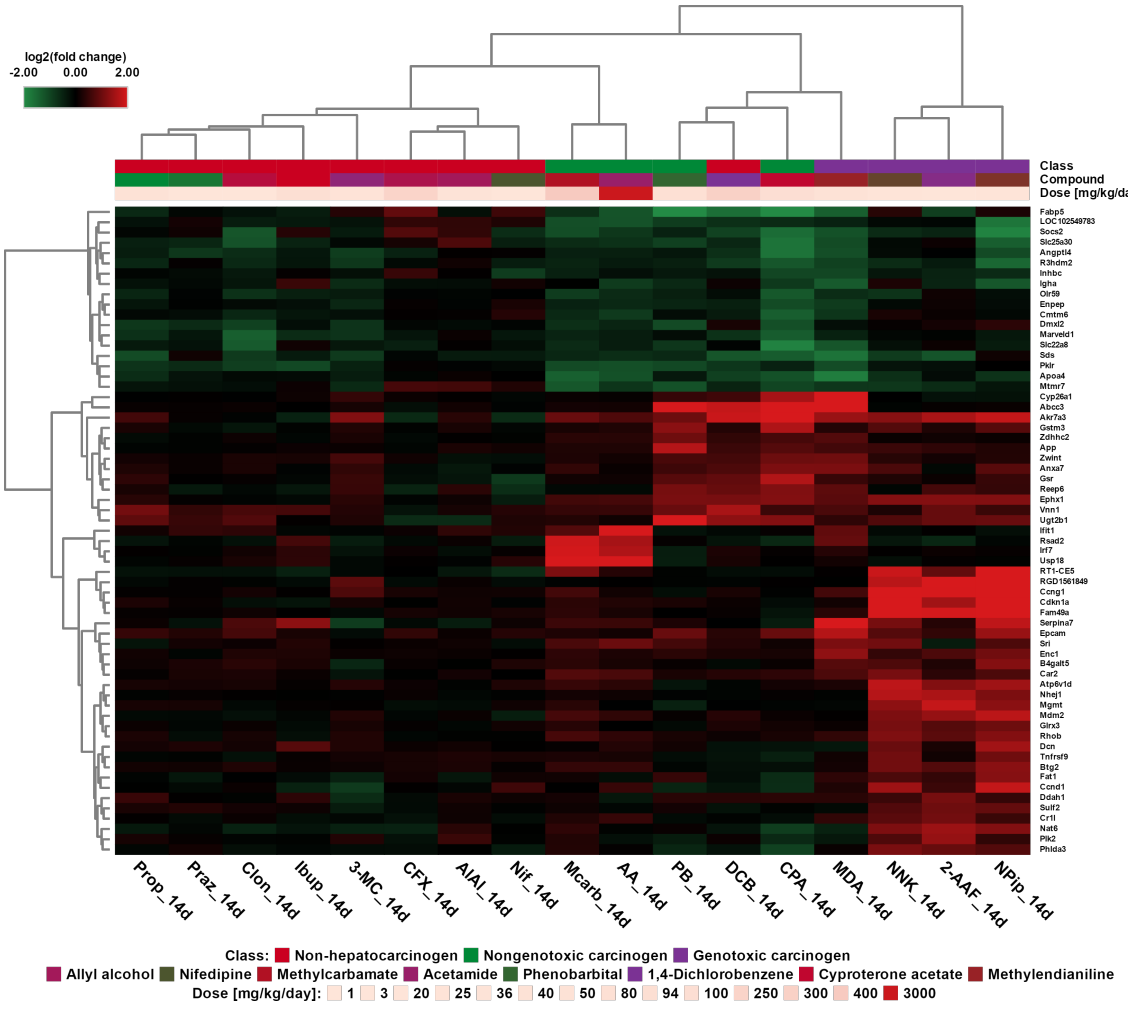


(b) Volcano plot

Figure 7.4: **Microarray data visualizations provided by MARCARviz.**



(c) Venn diagram



(d) Heat map

Figure 7.4: **Microarray data visualizations provided by MARCARviz (continued).**

Venn diagrams

Venn diagrams show the overlap of lists of deregulated genes for up to six different conditions, which enables the identification of genes that are deregulated in several conditions. They can be used to find genes that are affected by several nongenotoxic substances or by one substance at multiple time points. As with the other tools, the user can set the fold change and p-value thresholds used to identify deregulated genes or choose to include only up- or downregulated genes. The MARCARviz Venn diagrams also support cross-platform and cross-species comparisons by mapping genes to orthologous genes across species. The lists of shared deregulated genes can be saved for other analyses or downloaded as standard text files. Figure 7.4(c) shows a Venn diagram created with MARCARviz.

Heat maps

The most powerful tool for visual data exploration in MARCARviz are heat maps, i.e., color-coded matrix representations of the strength of gene deregulation in multiple conditions. This enables an easy, visual identification of common expression patterns that are shared across several conditions. Control conditions, e.g., NCs or GCs, can be included to visually identify genes that are deregulated specifically in the conditions of interest.

The visual exploration is further facilitated by the interactivity of the heat map: the user can zoom in on genes or gene clusters that are interesting, search for specific genes, or hide conditions that are not of interest. Genes are clustered hierarchically by their expression pattern in the selected conditions to group co-expressed genes. Again, users can set individual thresholds for fold changes and p-values to exclude genes that are not deregulated in any selected condition. They can also provide a list of genes that should be shown in the heat map, e.g., from a previous analysis of deregulated genes in a specific condition. The heat map can be downloaded as a PNG file for inclusion in manuscripts or presentations. The genes that the user identified can be saved as a gene list for further analysis. Figure 7.4(d) shows a heat map created with MARCARviz.

A modified variant, gene set enrichment heat maps, allows the visual exploration for gene set enrichments. Here, the significance of the hypergeometric test for gene set enrichment is color coded in the heat map. This exploratory analysis can help with the generation of new hypotheses on the mechanisms of NGCs, both on the level of single genes and gene sets.

7.3 Use case: Identification of phenobarbital target genes and pathways

Phenobarbital is an anticonvulsant drug that is used to treat many types of seizures in patients and has been in use for over a century. It is included in the WHO Model List of Essential Medicines. However, it has repeatedly been shown that phenobarbital acts as a nongenotoxic hepatocarcinogen in male and female mice (see, e.g., in the Carcinogenic Potency Database by Fitzpatrick, 2008) and is listed as a group 2B carcinogen by the IARC. Therefore, phenobarbital has been extensively studied as a model substance for NGCs.

Here, we used MARCARviz and a MARCAR dataset to identify potential target genes of phenobarbital, pathways that were affected by treatment with phenobarbital, and investigated the dependence of these effects on the constitutive androstane receptor (CAR) and the pregnane X receptor (PXR). The dataset is available from the Gene Expression Omnibus under the accession number GSE60684. An extensive analysis of this dataset was previously published by Luisier *et al.* (2014). This use case is also available as an interactive example online at the MARCARviz web platform.

The dataset was generated with Affymetrix microarrays. We assessed the array and sample quality, normalized the raw data with the RMA method, and calculated fold changes and Benjamini-Hochberg multiple-testing corrected p-values as described in Section 7.1.

As the first step in our analysis, we used a differential expression analysis to identify potential target genes. We selected genes that were at least two-fold up- or downregulated and showed significant differences between treated and control animals (corrected p-value ≤ 0.05) for wild-type mice receiving phenobarbital each day for up to 13 weeks. Gene expression has been profiled with microarrays at five time points, after 1, 7, 14, 28, and 91 days. At all five time points, *Cyp2b10* and *Cyp2c55* were among the top deregulated genes.

With a Venn diagram, we identified 11 genes that were significantly up- or downregulated at all five time points, among them four *Cyp* genes (*Cyp2b10*, *Cyp2c55*, *Cyp2c37*, and *Cyp2c54*), the *Wnt* signaling inhibitor *Wisp1*, and other genes that have previously been linked with phenobarbital treatment, e.g., *Gstm3* (Lempiäinen *et al.*, 2011; Lempiainen *et al.*, 2013). Similar observations were made in mice in which CAR and PXR have been replaced with humanized CAR and PXR. Again, *Cyp2b10* and *Cyp2c55* were among the top deregulated genes at all five time points, and five other genes were deregulated at all five time points: *Abcc4*, *Akr1b7*, *Cbr3*, *Gstm3*, and *Por*. In contrast, in mice with CAR and PXR knock outs, differential regulation was almost entirely eliminated at all time points.

We also performed gene set enrichment to find pathways that were deregulated after treatment with phenobarbital. For wild-type mice, the most affected KEGG pathways were the drug metabolism by cytochrome P450, glutathione metabolism, and retinol me-

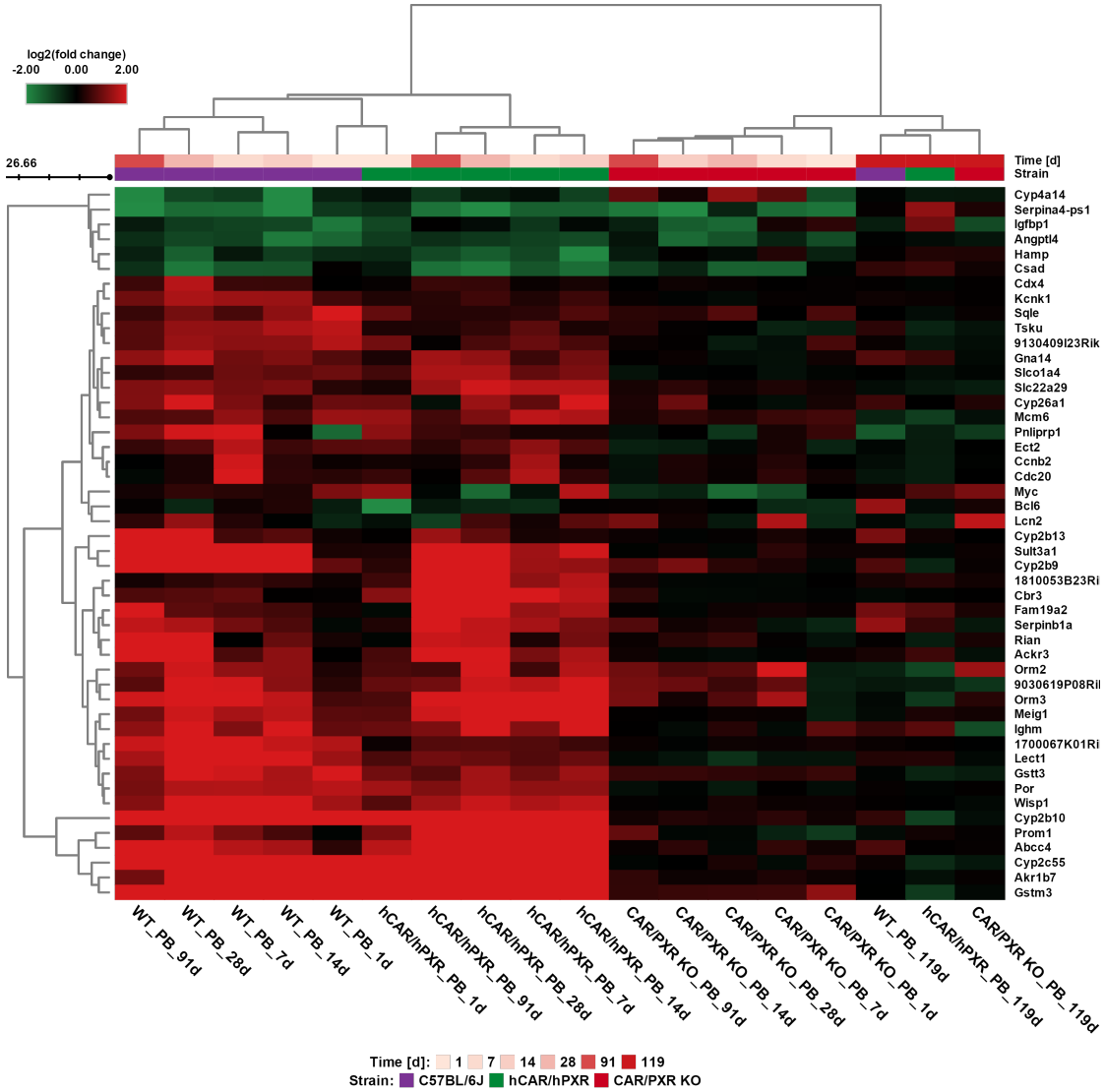


Figure 7.5: **Heat map of deregulated genes after treatment with phenobarbital of wild-type mice, mice with humanized CAR/PXR and CAR/PXR knockout mice.** Genes have been filtered for three-fold up- or downregulation and significant deregulation (corrected limma p-value ≤ 0.01) in at least one condition. Each row corresponds to a gene, each column to a condition (i.e., a combination of time point and mouse strain), each cell shows the $\log_2$ fold change between treated and vehicle control samples. The annotation bar above the heat map shows the time point and strain for each condition. The hierarchical clustering demonstrates the similarities of phenobarbital-mediated effects in the wild-type and humanized CAR/PXR mice. In contrast, gene deregulation is almost completely eliminated in CAR/PXR knockout mice. After recovery (conditions after 119 days), gene expression in wild-type and humanized CAR/PXR mice is similar to the knockout mice.

tabolism. These were consistently deregulated (q-value < 0.001) at all five time points. Again, we found very similar results for the mice with humanized CAR and PXR. For the CAR and PXR knockout mice, no pathway deregulation was observed as expected due to the lack of deregulation of individual genes. These results obtained with MARCARviz are in concordance with the results of the analysis performed by Luisier *et al.* (2014).

To visualize the major effects of phenobarbital on gene expression in wild type, humanized CAR/PXR, and CAR/PXR knockout mice, we created a heat map using the same filter criteria (two-fold deregulation and corrected p-value ≤ 0.05) with clustering of both genes and conditions, which is shown in Fig. 7.5. The clustering of the conditions shows a large difference in gene expression between wild-type and humanized CAR/PXR mice on one hand and the CAR/PXR knockout mice on the other, which confirms the analysis reported by Luisier *et al.* (2014).

7.4 Summary and conclusions

The MARCARviz web platform allows researchers to perform the most commonly used analyses and visualizations for microarray data. Summaries for genes and pathways are provided together with links to databases that contain additional information. This allows researchers to address a wide range of questions using the MARCAR data. In general, these can be classified into three overall tasks: identifying differentially regulated genes in a condition, mechanistic analysis of these genes, and comparison of the observed effects in two or more conditions. By facilitating these analyses, MARCARviz provides toxicological researchers with the opportunity to generate and confirm mechanistic hypotheses supported by the MARCAR data, which is one of the largest resources for molecular effects of short-term exposures of rodents to NGCs.

The interactive results can easily be shared with collaborators by sending links to the result web page. Alternatively, all tables and figures can be exported to standard table and image formats. The MARCAR data is additionally provided through GEO, such that bioinformatics analyses could also be performed starting with these raw data. However, MARCARviz provides preprocessed, ready-for-analysis data along with a user interface and no necessity to install any additional software. Only a browser with HTML and JavaScript support is required and all major browsers are supported.

The cross-species and cross-platform comparison of multiple datasets that is offered by MARCARviz is a particular advantage that would otherwise require mapping of the different manufacturer identifiers and identifying orthologous genes across species. We also provide all preprocessed data for download to allow researchers the use of their preferred tools to perform analyses that are not already offered by MARCARviz.

In conclusion, MARCARviz is a web platform for interactive visual exploration of the effects of NGCs on transcriptional regulation. We collected data from 16 datasets comprising 274 different conditions that were generated and analyzed over the course of the MARCAR project. The data included in MARCARviz cover the two most used rodent

species in preclinical risk assessment: mouse (*Mus musculus*) and rat (*Rattus norvegicus*), as well as human *in vitro* data. We provide the most commonly used analyses and visualizations for microarray data in an interactive fashion that allows visual exploration to aid the generation and validation of new hypotheses about the mechanism of NGCs. This will facilitate the discovery of biomarkers for NGC exposure and the development of new methods for early detection of carcinogenic effects in preclinical risk assessment.

MARCARviz is available from `https://tea.cs.uni-tuebingen.de/`. In addition, we provide the framework under MIT License on GitHub [7]. The framework can be used by other groups that use high-dimensional expression data to set up their custom web platforms for expression data analysis. There are no restrictions on its use by academic users.

[7] `https://github.com/mroemer/marcarviz`

Chapter 8

Summary and general conclusions

During its 5-year duration, the MARCAR project has generated a rich set of toxicogenomics data to study the effects of NGCs. Ultimately, the goal of the MARCAR project was to provide a proof of concept that early molecular biomarkers can be used to predict later cancer development. To investigate the effects of NGCs on the molecular level, the MARCAR consortium used several high-throughput technologies that interrogate the transcriptome, proteome, and methylome. The volume of the generated data requires bioinformatics methods for efficient processing, exploration, and visualization. To facilitate the access to bioinformatics software for life science researchers, this thesis provided the ZBIT Bioinformatics Toolbox, a set of bioinformatics tools for the analysis of biological data that can be applied in several biological fields (Chapter 4). We also developed a new similarity scoring approach for gene expression profiles that can be used to identify substances that have similar effects on the gene regulation in rodents (Chapter 5). As a proof of concept that the integration of early biomarkers from multiple *omics* platforms enables the reliable identification of nongenotoxic carcinogens (NGCs), we performed a series of experiments with a multi-*omics* dataset generated in the MARCAR project (Chapter 6). Finally, we designed and implemented a web platform that allows other researchers to inspect the MARCAR gene expression datasets interactively and visually (Chapter 7). This chapter summarizes the four individual contributions and provides a general conclusion of this thesis.

The ZBIT Bioinformatics Toolbox

Chapter 4 describes the ZBIT Bioinformatics Toolbox (Römer *et al.*, 2016b), which is a web platform that provides bioinformatics tools for life science researchers. The rise of high-throughput technologies and the increase in computational resources enabled many new insights and gave birth to bioinformatics, which encompasses disciplines on the border of biology and computer science, such as systems biology, genomics, transcriptomics, and a whole range of other *omics* disciplines. However, life science researchers often struggle with the required technical knowledge that is needed to operate bioinformatics software. Currently, bioinformatics software often depends on specific operating systems, third-party libraries, or command-line interface skills. To facilitate the access to a range of bioinformatics tools that were developed by our bioinformatics team, we

implemented a web platform that enables life science researchers to use our tools without the problems mentioned above. Users can execute all tools from any web browser without installing software on their local machine and independent of their operating system or hardware configuration because all analyses are run on our computation cluster.

The ZBIT Bioinformatics Toolbox encompasses eight tools from three important fields of bioinformatics: systems biology, transcription factor analysis, and expression data analysis. In systems biology, we provide tools for processing and annotating models in the systems biology community standards BioPAX and SBML. BioPAX2SBML translates models from BioPAX into SBML and maintains qualitative information. SBMLsqueezer adds kinetic rate law equations to qualitative SBML models to enable model simulation. SBML2LaTeX generates human-readable reports for model checking and exchange. ModelPolisher enriches SBML models with information from the BiGG database. In transcription factor annotation, we provide the tools TFpredict and SABINE, which can be used to identify and annotate transcription factors from the protein sequence and predict the DNA motif that is bound by the transcription factor. For expression data analysis, we provide RPPApipe for the processing, analysis, and visualization of RPPA data and ToxDBScan to identify chemicals that induce similar gene expression changes in rodents after subchronic treatment.

The ZBIT Bioinformatics Toolbox uses the Galaxy framework, which is well established in the life science community and offers a familiar interface for researchers. Also, by using the Galaxy framework, the ZBIT Bioinformatics Toolbox provides job histories, data management, and workflow editors which adhere to scientific standards for reproducibility and the storage of results. We have compared the ZBIT Bioinformatics Toolbox with competing software and web platforms in the target bioinformatics areas and demonstrated that our tools and the ZBIT Bioinformatics Toolbox provide significant advantages over existing software. Furthermore, we recorded usage statistics to evaluate if and how our web platform is frequented by users. In the one-year period from May 2015 to June 2016, we recorded 4,403 individual jobs that were submitted by 16 registered users and an unknown number of anonymous users. The most used tool was RPPApipe, with over 2,000 individual tool executions. In one case that we know of, the ZBIT Bioinformatics Toolbox inspired the integration of RPPApipe into the standard workflow of a research group.

In summary, the ZBIT Bioinformatics Toolbox provides life science researchers with an easier method of using bioinformatics software. Because no software must be installed locally and jobs are executed on our computation cluster, researchers can use bioinformatics on any device, e.g., low-end desktop computers or mobile devices in the lab. Also, the ZBIT Bioinformatics Toolbox offers an opportunity to evaluate tools before installing them locally for integration with other tools. Based on user feedback and usage statistics, we have improved and extended the ZBIT Bioinformatics Toolbox and will continue to include new tools as they are developed, as demonstrated by the addition of ModelPolisher earlier this year.

ToxDBScan

Chapter 5 introduces a novel similarity scoring method for gene expression profiles (Römer *et al.*, 2014b). Microarrays and more recently RNAseq allow the large-scale profiling of gene expression, which has become a standard method in many areas of biology, including toxicogenomics. However, the comparison of gene expression profiles from experiments run at different times or in different labs remains challenging due to systematic biases that introduce batch effects and affect the dynamic range of the measured expression. Most existing similarity measures for gene expression profiles (e.g., Pearson correlation or Euclidean distance) depend on the strength of the observed effects, which can confound analyses when the dynamic range differs between experiments.

We developed a new similarity measure that is based on chemoinformatics concepts and set theory and uses only information on up- and downregulation of genes without considering the strength of the regulation. The sets of up- and downregulated genes define the gene expression fingerprint of an experiment, for which the similarity with fingerprints of other experiments can be calculated. To this end, we developed a modified Jaccard index based on the idea of the Tanimoto index. In short, the genes that are deregulated in the same direction (up or down) in both experiments are compared to the total number of observed genes to determine the similarity. Furthermore, we introduced a weighting of each gene according to the information content concept from information theory. A gene is considered to have higher information content, and thus a higher weight if it is regulated in few experiments, whereas genes that are deregulated in many experiments have a lower information content and thus lower weight.

To evaluate our new similarity scoring method, we performed a series of experiments with toxicogenomics datasets. First, we obtained two large toxicogenomics datasets (TG-GATEs and DrugMatrix) to construct a reference database of gene expression fingerprints and calculated the information content of the genes for which gene expression was measured. Next, we calculated gene expression fingerprints of an independent dataset from the MARCAR project that used a similar experimental setting but other dosages and time points than TG-GATEs and DrugMatrix. Using the fingerprints of several compounds used in the MARCAR experiment, we assessed whether the ten compounds that were present in the reference database and the MARCAR experiment could be recalled from the reference database with the MARCAR fingerprints. All ten substances were successfully recalled as either the most or the second-most similar compound from the over 200 compounds in the reference database. For these ten substances and the five substances only present in the MARCAR experiment, we also observed that compounds with a similar biological MOA were enriched in fingerprints with high relative similarity (≥ 0.8). These results with independent data indicate that our similarity measure can successfully detect similarities in gene expression experiments.

Next, we evaluated if our similarity scoring method can be used to identify NGCs. We used publicly available carcinogenicity annotations for the compounds in the three datasets. The MARCAR dataset was used as the evaluation dataset and TG-GATEs and

DrugMatrix to construct the reference database. All 15 substances in the MARCAR dataset were correctly classified as either NGC, GC, or NC.

To make our reference database and the similarity scoring available to the public, we developed ToxDBScan. After performing gene expression profiling, toxicologists can use ToxDBScan to identify substances that induce similar gene expression profiles. This can provide leads for the mechanistic analysis, e.g., to determine the MOA if tumor growth is observed in an experiment. ToxDBScan also offers visualizations and pathway enrichment analysis to facilitate the interpretation of gene expression profiles further. The major advantage of both ToxDBScan and the underlying similarity scoring approach is the platform independence, i.e., users can upload gene expression fingerprints that have been obtained with any expression profiling technology. ToxDBScan is available as a web tool from the ZBIT Bioinformatics Toolbox.

Multi-*omics* prediction of nongenotoxic carcinogenicity

Chapter 6 demonstrates that the integration of data from multiple *omics* platforms improves the identification of NGCs (Römer *et al.*, 2014a). Predictive toxicogenomics studies have shown that gene expression profiling can be used to find early biomarkers that can identify nongenotoxic substances significantly earlier than the traditional LRB. However, regulatory agencies have not yet implemented *omics*-based short-term assays for NGCs into the preclinical development of new drugs due to the lack of mechanistic understanding and low overlap in the biomarkers identified in different studies. Currently, most toxicogenomics studies use only mRNA microarrays to search for biomarkers. Only very few have explored other platforms, such as miRNA microarrays or protein arrays, but did not attempt to integrate these with gene expression data.

We explored the use of multiple *omics* platforms to identify changes that are conserved across several layers of gene regulation. Furthermore, we developed two new feature representations for multi-*omics* data which integrate the expression observed across several platforms into a single value. The pathway enrichment features use information from pathway databases to perform an integrated pathway enrichment that assesses the multi-layer deregulation of biological pathways. The molecular interaction features use databases of gene-regulatory interactions to identify changes in multiple, connected levels of regulation.

To evaluate the multi-platform approach and our new features, we collected a new short-term toxicogenomics dataset that encompasses mRNA, miRNA, and protein expression data for a set of NCs, NGCs, and GCs. Using the new dataset, we trained predictive models on individual platform data, multi-platform data, and multi-platform data enriched with integrative features. The model training and evaluation was performed in a 2×2-nested CV (with parameter optimization) that was replicated ten times to get an unbiased estimate of the classification performance. We performed the evaluation for three different contrasts: C vs. NC, GC vs. NGC, and NGC vs. NC. Across all three contrasts we observed an increase in performance when we used the combined multi-

platform features compared to each individual platform. The multi-platform prediction achieved AUCs > 0.9 in all three contrasts and > 0.97 for C vs. NC and NGC vs. NC. Except for the protein signature in the NGC vs. GC contrast, the multi-platform signature performed better than the individual signatures in all settings. By additionally using the new pathway enrichment and molecular interaction features, the prediction performance was further increased. This demonstrates that our integrative approach can improve the identification of (nongenotoxic) carcinogens in short-term toxicogenomics studies.

The integrative pathway enrichment and molecular interaction features offer an abstraction of molecular changes to molecular interactions and biological pathways. This abstraction can help researchers identify relevant systemic changes that may not be apparent from the individual expression profiles, e.g., if compounds act by altering the same downstream pathway by perturbing the regulation of different genes or miRNAs. Also, we expect that changes that are conserved across multiple layers of gene regulation will provide more robust biomarkers. Our abstract feature representations incorporate this multi-layer aspect directly into the model training.

In summary, we investigated the use of multiple *omics* platforms for biomarker discovery in toxicogenomics. We showed that the classification accuracy of models trained with data from multiple *omics* platforms is higher than for models trained on data from the individual platforms. Also, we developed two integrative feature representations for multi-*omics* data that further increase the performance in combination with multi-platform data. Although the number of compounds in our study was very small, we demonstrated that toxicogenomics can benefit from the integration of data from multiple *omics* platforms.

MARCARviz

Chapter 7 describes MARCARviz, a web platform that provides tools for the interactive analysis and visualization of the MARCAR transcriptomics data. As discussed in Chapter 6, high-throughput mRNA expression profiling is the de facto standard in toxicogenomics studies for biomarker discovery. However, toxicologists often require the help of bioinformaticians to analyze and visualize the high-dimensional expression data that is produced by modern mRNA profiling techniques. While raw and normalized expression data is often made available to the public through platforms such as ArrayExpress or the Gene Expression Omnibus, researchers often lack the resources or the bioinformatics support to analyze public data. Using modern web technologies, we simplified the analysis and visualization of the MARCAR expression data for researchers without a technical background. To this end, we implemented MARCARviz, a web platform that is accessible through the web browser and offers the most used analyses and visualizations for mRNA expression data, e.g., heat maps, differential expression analysis, and pathway enrichment. MARCARviz provides the preprocessed mRNA expression data of all MARCAR studies together with meta-information on the conditions that were used in each study, quality control results, and additional information on genes and pathways

from external databases.

The architecture of the MARCARviz web platform consists of three layers: the front-end that is visible to the user, the back-end that stores and analyzes the expression data, and the middleware, which handles the communication between front-end and back-end. The MARCARviz front-end uses JavaScript visualization libraries and jQuery plugins to render expression tables and plots directly in the user's web browser. Therefore, users require only a web browser to analyze MARCAR data and need not install any additional software or download large amounts of raw expression data. As middleware, we use a Node.js server that receives the analysis requests that are submitted by users through the front-end. These requests are delegated to a dedicated, internal server that hosts the back-end. The back-end uses a collection of R scripts to process the data, format the results according to the requirements of the front-end JavaScript libraries, and retrieves the expression data from a MongoDB database. When the analysis is finished, the results are passed back by the Node.js server and rendered for the user by the front-end.

MARCARviz improves the visibility of the MARCAR expression data by providing interactive analysis and visualization tools for researchers without strong technical background. The framework that was used to set up MARCARviz is provided as a proof-of-concept for interactive data analysis of expression data in the web browser. Thus, other groups can use the framework to provide custom web platforms for their expression data collections. We hope that this improves the efficiency of the current data sharing, which is hindered by the need for technical and bioinformatics support to analyze external data. We think that this can lead to a better leveraging of the existing large amounts of expression data that are available as raw data from ArrayExpress or the Gene Expression Omnibus, which, in turn, will provide more biological insights and reduce the number of further animal experiments.

General conclusions and outlook

In the framework of the MARCAR project, this thesis explored novel approaches and developed new tools for the detection and characterization of NGCs during preclinical risk assessment. The focus of this thesis was on machine learning techniques for cancer risk prediction and the visualization of toxicogenomics data. We described four individual contributions, which provide bioinformaticians and toxicologists with new tools for the analysis of toxicogenomics data. The developed tools are web-based in order to be accessible and easily usable for users. We also looked into the integration of gene expression data (transcriptomics) with other *omics* data, e.g., from proteomics. We expect that future studies can build on the developed tools and methods to confirm and extend upon the results described in this thesis.

Appendix A

Supplementary Tables

Table A.1: **Predefined workflows in the ZBIT Bioinformatics Toolbox.**

Workflow name	Description	Steps
BioPAX2SBML and Squeeze2LaTeX	Converts BioPAX files to full SBML models and human-readable reports.	3
TFpredict & SABINE	Uses TFpredict and SABINE to annotate transcription factors.	2
RPPApipe two-class	Processes datasets with paired samples obtained from RPPAs.	12
RPPApipe time-series	Processes replicated time-series datasets obtained from RPPAs.	12
RPPApipe multi-class	Processes datasets with multiple sample classes obtained from RPPAs.	12

Table A.2: **Chemicals used for evaluation of the similarity scoring index.** Male Wistar rats were treated with the chemicals each day for up to 14 days. For each chemical, the Chemical Abstracts Service (CAS) registry number, dosing time and dose is listed, as well as the short name that is used in the tables and figures. The last column lists the databases that contain the test compound (DM = DrugMatrix, TGG = TG-GATEs). Table adapted from Römer *et al.* (Römer *et al.*, 2014a).

Compound	Short Name	CAS Number	Dosing Time (day)	Dose (mg/kg/day)	Contained in
Genotoxic carcinogens (GCs)					
Direct Black 38	CIDB	1937-37-7	7	146	-
Nitrosodimethylamine	DMN	62-75-9	7	4	DM
Nongenotoxic carcinogens (NGCs)					
Piperonyl butoxide	PBO	51-03-6	3	1200	-
Methyl carbamate	MCA	598-55-0	14	400	-
Dehydroepiandrosterone	DHEA	53-43-0	14	600	-
Methapyrilene	MP	135-23-9	14	60	TGG, DM
Thioacetamide	TAA	62-55-5	7	19.2	TGG, DM
Diethylstilbestrol	DES	56-53-1	3	10	DM
Wy-14643	WY	50892-23-4	3	60	TGG, DM
Acetamide	AAA	60-35-5	14	3000	TGG
Ethionine	ET	67-21-0	14	200	TGG
Cyproterone acetate	CPR	427-51-0	14	100	DM
Phenobarbital	PB	50-06-6	14	80	TGG, DM
Non-hepatocarcinogens (NCs)					
Cefuroxime	CFX	55268-75-2	14	250	
Nifedipine	NIF	21829-25-4	14	3	TGG

Table A.3: **Compounds used in subacute studies that were performed during the MARCAR project.** The compounds were manually selected and classified by the MARCAR consortium according to their carcinogenic effects in the liver of rats and mice.

Classification	Abbreviation	Compound	CAS
Genotoxic carcinogen	2-AAF	2-Acetylaminofluorene	53-96-3
	2-NF	2-Nitrofluorene	607-57-8
	AB1	Aflatoxin B1	1162-65-8
	CIDB	C.I Direct Black	1937-37-7
	DMN	Dimethylnitrosamine	56-23-5
	MDA	Methylendianiline	101-77-9
	NNK	4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone	64091-91-4
	NNM	N-Nitrososmorpholine	59-89-2
	NPip	N-Nitrosopiperidine	100-75-4
Nongenotoxic carcinogen	AA	Acetamide	60-35-5
	Aap	Acetaminophen	103-90-2
	CITCO	CITCO	338404-52-7
	CPA	Cyproterone acetate	427-51-0
	DES	Diethyl-stilbestrol	56-53-1
	DHEA	Dehydroepiandrosterone	53-43-0
	ETH	Ethionine	67-21-0
	Mcarb	Methylcarbamate	589-55-0
	MPy	Methapyrilene HCl	135-23-9
	PB	Phenobarbital	50-06-6
	PBO	Piperonyl-butoxide	51-03-6
	TAA	Thioacetamid	62-55-5
	Wy	Wy-14643	50892-23-4
Noncarcinogen	3-MC	3-Methylcholanthrene	56-49-5
	AlAl	Allyl alcohol	107-18-6
	CFX	Cefuroxime	55268-75-2
	Clon	Clonidine	4205-90-7
	DCB	1,4-Dichlorobenzene	106-46-7
	Ibup	Ibuprofen	15687-27-1
	Nif	Nifedipine	21829-25-4
	Pio	Pioglitazone	111025-46-8
	Praz	Prazosin	19216-56-9
	Prop	Propranolol	525-66-6
	Rosi	Rosiglitazone	122320-73-4

Table A.4: **Datasets generated in the MARCAR project and available in MARCARviz.** All datasets listed in this table are part of the MARCAR Gene Expression Omnibus SuperSeries "IMI MARCAR Project: towards novel biomarkers for cancer risk assessment", which is available from Gene Expression Omnibus under the accession number GSE68387.

GEO	Summary	Species	Microarray
GSE68969	Effect of PB in APC-KO mice	Mouse	Mouse Genome 430 2.0
GSE68779	Effects of PB in Ctnnb1-KO mice	Mouse	Mouse Genome 430 2.0
GSE68592	Effects of Pio and Rosi in rat bladder	Rat	Rat Gene Expression 8x60K G4853A
GSE68365	CAR and PXR dependent PB effects in C57BL mice	Mouse	Mouse Genome 430 2.0
GSE68364	Effects of NGC in C57BL mice after 4 weeks	Mouse	Mouse Genome 430 2.0
GSE68361	Effects of PB in CD1 mice after 4 and 13 weeks	Mouse	Mouse Genome 430 2.0
GSE68128	Effects of NGCs in Wistar rats after 4 and 13 weeks	Rat	Rat Gene ST 2.0
GSE68121	Effects of PB on tumorigenesis in rats	Rat	Rat Genome 230 2.0
GSE68120	Effects of PB and CPA in HCs and MCs of rats	Rat	Rat Genome 230 2.0
GSE68111	Effects of PB in HCs and MCs in mice	Mouse	Mouse Genome 430A 2.0
GSE68110	Effects of NGCs in Wistar rats after 2 weeks	Rat	Rat Expression Array 230A
GSE60684	CAR- and PXR dependent effects of PB in mice	Mouse	Mouse Genome 430 2.0
GSE51355	Ha-ras and *beta*-catenin mut. mice tumors	Mouse	Mouse Genome 430 2.0
GSE44783	Effects of NGCs in CD1 mice	Mouse	Mouse Genome 430 2.0
GSE34423	Effects of PB in B6C3F1 mice	Mouse	Mouse Genome 430 2.0
GSE68493	CAR/PXR dependent effects of PB in human hepatocytes	Human	Human Genome U133 Plus 2.0

Abbreviations

3'-UTR	3' untranslated region
ACC	accuracy
AUC	area under the ROC curve
BGLM	Bayesian Generalized Linear Models
BioPAX	Biological Pathway Exchange
C	carcinogen
CAS	Chemical Abstracts Service
CPDB	Carcinogenic Potency Database
CV	cross-validation
DBD	DNA-binding domain
DEG	differentially expressed gene
DNA	deoxyribonucleic acid
GC	genotoxic carcinogen
GEO	Gene Expression Omnibus
LOO	leave-one-out validation
LRB	lifetime rodent cancer bioassay
miRNA	microRNA
MLP	multi-layer perceptron
MM probe	mismatch probe
MOA	mode of action
mRNA	messenger RNA
NC	non-hepatocarcinogen
NGC	nongenotoxic carcinogen
NN	Neural Network
nt	nucleotide
PCA	principal components analysis
PCR	Principal Component Regression
PFM	position frequency matrix
PID	Pathway Interaction Database
PM probe	perfect match probe
PTM	post-translational modification
RF	Random Forest
RNA	ribonucleic acid
RMA	Robust multi-array average

RPPA	reverse phase protein array
SABINE	Stand-alone binding specificity estimator
SBGN	Systems Biology Graphical Notation
SBML	Systems Biology Markup Language
SBO	Systems Biology Ontology
SVM	Support Vector Machine
TF	transcription factor
TG-GATEs	Toxicogenomics Project-Genome Assisted Toxicity Evaluation System
XML	Extensible Markup Language

Bibliography

Afshari, C. A., Hamadeh, H. K., and Bushel, P. R. (2011). The Evolution of Bioinformatics in Toxicology: Advancing Toxicogenomics. *Toxicological Sciences*, **120**(S1), S225–S237.

Airola, A., Pahikkala, T., Waegeman, W., De Baets, B., and Salakoski, T. (2009). A comparison of AUC estimators in small-sample studies. In *Proceedings of the 3rd International workshop on Machine Learning in Systems Biology*, pages 15—-23.

Allfrey, V. G., Faulkner, R., and Mirsky, A. E. (1964). Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proceedings of the National Academy of Sciences*, **51**, 786–94.

Alwine, J. C., Kemp, D. J., and Stark, G. R. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences*, **74**(12), 5350–5354.

Amaratunga, D. and Cabrera, J. (2001). Analysis of Data From Viral DNA Microchips. *Journal of the American Statistical Association*, **96**(456), 1161–1170.

Ames, B. N. and Gold, L. S. (1990). Chemical carcinogenesis: too many rodent carcinogens. *Proceedings of the National Academy of Sciences*, **87**(19), 7772–7776.

Ames, B. N., McCann, J., and Yamasaki, E. (1975). Methods for detecting carcinogens and mutagens with the salmonella/mammalian-microsome mutagenicity test. *Mutation Research*, **31**(6), 347–363.

Anand, P., Kunnumakara, A. B., Sundaram, C., Harikumar, K. B., Tharakan, S. T., Lai, O. S., Sung, B., and Aggarwal, B. B. (2008). Cancer is a Preventable Disease that Requires Major Lifestyle Changes. *Pharmaceutical Research*, **25**(9), 2097–2116.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., *et al.* (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**(1), 25–29.

Auerbach, S. S., Shah, R. R., Mav, D., Smith, C. S., Walker, N. J., Vallant, M. K., Boorman, G. A., and Irwin, R. D. (2010). Predicting the hepatocarcinogenic potential of

alkenylbenzene flavoring agents using toxicogenomics and machine learning. *Toxicology and Applied Pharmacology*, **243**(3), 300–314.

Bardou, P., Mariette, J., Escudié, F., Djemiel, C., and Klopp, C. (2014). jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics*, **15**(1), 293.

Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W.-C., Ledoux, P., Rudnev, D., Lash, A. E., Fujibuchi, W., and Edgar, R. (2004). NCBI GEO: mining millions of expression profiles–database and tools. *Nucleic Acids Research*, **33**(D1), D562–D566.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., *et al.* (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Research*, **41**(D1), D991–D995.

Bartel, D. P. (2004). MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell*, **116**(2), 281–297.

Bartel, D. P. (2009). MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, **136**(2), 215–233.

Bartel, D. P. and Chen, C.-Z. (2004). Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. *Nature Reviews Genetics*, **5**(5), 396–400.

Becich, M. (2000). The role of the pathologist as tissue refiner and data miner: The impact of functional genomics on the modern pathology laboratory and the critical roles of pathology informatics and bioinformatics. *Molecular Diagnosis*, **5**(4), 287–299.

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., and Rätsch, G. (2008). Support Vector Machines and Kernels for Computational Biology. *PLoS Computational Biology*, **4**(10), e1000173.

Benigni, R., Bossa, C., and Tcheremenskaia, O. (2013). Nongenotoxic Carcinogenicity of Chemicals: Mechanisms of Action and Early Recognition through a New Set of Structural Alerts. *Chemical Reviews*, **113**(5), 2940–2957.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, **57**(1), 289–300.

Bernstein, C., Prasad, A. R., Nfonsam, V., and Bernstein, H. (2013). DNA Damage, DNA Repair and Cancer. In C. Chen, editor, *New Research Directions in DNA Repair*. InTech.

Biggar, K. K. and Li, S. S.-C. (2014). Non-histone protein methylation as a regulator of cellular signalling and function. *Nature Reviews Molecular Cell Biology*, **16**(1), 5–17.

Boffetta, P. and Hashibe, M. (2006). Alcohol and cancer. *The Lancet Oncology*, **7**(2), 149–156.

Bolstad, B. (2004). *Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization*. Ph.D. thesis, University of California, Berkeley.

Bolstad, B., Irizarry, R., Astrand, M., and Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**(2), 185–193.

Bolten, B. M. and DeGregorio, T. (2002). Trends in development cycles. *Nature Reviews Drug Discovery*, **1**(5), 335–336.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*, pages 144–152, New York, New York, USA. ACM Press.

Braeuning, A., Singh, Y., Rignall, B., Buchmann, A., Hammad, S., Othman, A., Recklinghausen, I., Godoy, P., *et al.* (2010). Phenotype and growth behavior of residual β-catenin-positive hepatocytes in livers of β-catenin-deficient mice. *Histochemistry and Cell Biology*, **134**(5), 469–481.

Braeuning, A., Gavrilov, A., Geissler, M., Wenz, C., Colnot, S., Templin, M. F., Metzger, U., Römer, M., Zell, A., and Schwarz, M. (2016). Tumor promotion and inhibition by phenobarbital in livers of conditional Apc-deficient mice. *Archives of Toxicology*, pages 1–14.

Brazma, A. and Vilo, J. (2000). Gene expression data analysis. *FEBS Letters*, **480**(1), 17–24.

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**(1), 5–32.

Büchel, F., Wrzodek, C., Mittag, F., Dräger, A., Eichner, J., Rodriguez, N., Le Novère, N., and Zell, A. (2012). Qualitative translation of relations from BioPAX to SBML qual. *Bioinformatics*, **28**(20), 2648–2653.

Büchel, F., Rodriguez, N., Swainston, N., Wrzodek, C., Czauderna, T., Keller, R., Mittag, F., Schubert, M., *et al.* (2013). Path2Models: large-scale generation of computational models from biochemical pathway maps. *BMC Systems Biology*, **7**(116).

Budunova, I. V. and Williams, G. M. (1994). Cell culture assays for chemicals with tumor-promoting or tumor-inhibiting activity based on the modulation of intercellular communication. *Cell Biology and Toxicology*, **10**(2), 71–116.

Bumgarner, R. (2013). Overview of DNA Microarrays: Types, Applications, and Their Future. In *Current Protocols in Molecular Biology*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Buness, A., Huber, W., Steiner, K., S ltmann, H., and Poustka, A. (2005). arrayMagic: two-colour cDNA microarray quality control and preprocessing. *Bioinformatics*, **21**(4), 554–556.

Cai, Y.-D., Liu, X.-J., Xu, X.-b., and Chou, K.-C. (2002). Prediction of protein structural classes by support vector machines. *Computers & Chemistry*, **26**(3), 293–296.

Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M., and Werner, T. (2005). MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, **21**(13), 2933–2942.

Cattley, R. C. and Popp, J. A. (1989). Differences between the promoting activities of the peroxisome proliferator WY-14,643 and phenobarbital in rat liver. *Cancer Research*, **49**(12), 3246–3251.

Cawley, G. C. and Talbot, N. L. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *The Journal of Machine Learning Research*, **11**, 2079–2107.

Chaouiya, C., Bérenguier, D., Keating, S. M., Naldi, A., van Iersel, M. P., Rodriguez, N., Dräger, A., Büchel, F., *et al.* (2013). SBML qualitative models: a model representation format and infrastructure to foster interactions between qualitative modelling formalisms and tools. *BMC Systems Biology*, **7**(135).

Chelliah, V., Juty, N., Ajmera, I., Ali, R., Dumousseau, M., Glont, M., Hucka, M., Jalowicki, G., *et al.* (2015). BioModels: ten-year anniversary. *Nucleic Acids Research*, **43**(D1), D542–D548.

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., and Liu, C. (2011). Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. *PLoS ONE*, **6**(2), e17238.

Chen, G. G., Zeng, Q., and Tse, G. M. (2008). Estrogen and its receptors in cancer. *Medicinal Research Reviews*, **28**(6), 954–974.

Chieli, E., Aliboni, F., Saviozzi, M., and Malvaldi, G. (1987). Induction of micronucleated erythrocytes by primary thioamides and their metabolites in the mouse. *Mutation Research*, **192**(2), 141–143.

Choudhary, C., Kumar, C., Gnad, F., Nielsen, M. L., Rehman, M., Walther, T. C., Olsen, J. V., and Mann, M. (2009). Lysine Acetylation Targets Protein Complexes and Co-Regulates Major Cellular Functions. *Science*, **325**(5942), 834–840.

Cohen, S. (1995). Human Relevance of Animal Carcinogenicity Studies. *Regulatory Toxicology and Pharmacology*, **21**(1), 75–80.

Cohen, S. M. (2010). Evaluation of Possible Carcinogenic Risk to Humans Based on Liver Tumors in Rodent Assays: The Two-Year Bioassay Is No Longer Necessary. *Toxicologic Pathology*, **38**(3), 487–501.

Cohen, S. M., Arnold, L. L., Eldan, M., Lewis, A. S., and Beck, B. D. (2006). Methylated Arsenicals: The Implications of Metabolism and Carcinogenicity Studies in Rodents to Human Risk Assessment. *Critical Reviews in Toxicology*, **36**(2), 99–133.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20**(3), 273–297.

Courtot, M., Juty, N., Knupfer, C., Waltemath, D., Zhukova, A., Drager, A., Dumontier, M., Finney, A., *et al.* (2011). Controlled vocabularies and semantics in systems biology. *Molecular Systems Biology*, **7**, 543.

Cover, T. and Thomas, J. (2006). *Elements of Information Theory 2nd Edition*. Wiley-Interscience.

Cowles, C., Mally, A., and Chipman, J. (2007). Different mechanisms of modulation of gap junction communication by non-genotoxic carcinogens in rat liver in vivo. *Toxicology*, **238**(1), 49–59.

Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., Bunney, W. E., Myers, R. M., *et al.* (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Research*, **33**(20), e175.

D'Andrade, P. N. and Fulmer-Smentek, S. (2012). Agilent MicroRNA Microarray Profiling System. In J.-B. Fan, editor, *Next-Generation MicroRNA Expression Profiling Technology*, pages 85–102. Humana Press, Totowa, NJ.

Darnell, J. E. (2002). Transcription factors as targets for cancer therapy. *Nature Reviews Cancer*, **2**(10), 740–749.

DeBerardinis, R. J. and Cheng, T. (2010). Q's next: the diverse functions of glutamine in metabolism, cell biology and cancer. *Oncogene*, **29**(3), 313–324.

Debouck, C. and Goodfellow, P. N. (1999). DNA microarrays in drug discovery and development. *Nature Genetics*, **21**, 48–50.

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

Deutsch, W. A., Kukreja, A., Shane, B., and Hegde, V. (2001). Phenobarbital, oxazepam and Wyeth 14,643 cause DNA damage as measured by the Comet assay. *Mutagenesis*, **16**(5), 439–442.

Dewa, Y., Nishimura, J., Muguruma, M., Jin, M., Kawai, M., Saegusa, Y., Okamura, T., Umemura, T., and Mitsumori, K. (2009). Involvement of oxidative stress in hepatocellular tumor-promoting activity of oxfendazole in rats. *Archives of Toxicology*, **83**(5), 503–511.

Dhami, G. K., Liu, H., Galka, M., Voss, C., Wei, R., Muranko, K., Kaneko, T., Cregan, S. P., Li, L., and Li, S. S.-C. (2013). Dynamic Methylation of Numb by Set8 Regulates Its Binding to p53 and Apoptosis. *Molecular Cell*, **50**(4), 565–576.

Dietrich, D. and Swenberg, J. (1991). Preneoplastic lesions in rodent kidney induced spontaneously or by non-genotoxic agents: predictive nature and comparison to lesions induced by genotoxic carcinogens. *Mutation Research*, **248**(2), 239–260.

DiMasi, J. A., Grabowski, H. G., and Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics*, **47**, 20–33.

Dogrusoz, U., Erson, E. Z., Giral, E., Demir, E., Babur, O., Cetintas, A., and Colak, R. (2006). PATIKAweb: a Web interface for analyzing biological pathways through advanced querying and visualization. *Bioinformatics*, **22**(3), 374–375.

Doktorova, T. Y., Yildirimman, R., Vinken, M., Vilardell, M., Vanhaecke, T., Gmuender, H., Bort, R., Brolen, G., *et al.* (2013). Transcriptomic responses generated by hepatocarcinogens in a battery of liver-based in vitro models. *Carcinogenesis*, **34**(6), 1393–402.

Dräger, A. and Palsson, B. Ø. (2014). Improving Collaboration by Standardization Efforts in Systems Biology. *Frontiers in Bioengineering and Biotechnology*, **2**, 61.

Dräger, A., Hassis, N., Supper, J., Schröder, A., and Zell, A. (2008). SBMLsqueezer: a CellDesigner plug-in to generate kinetic rate equations for biochemical networks. *BMC Systems Biology*, **2**(39).

Dräger, A., Planatscher, H., Motsou Wouamba, D., Schröder, A., Hucka, M., Endler, L., Golebiewski, M., Müller, W., and Zell, A. (2009). SBML2L(A)T(E)X: conversion of SBML files into human-readable reports. *Bioinformatics*, **25**(11), 1455–1456.

Dunning, M. J., Smith, M. L., Ritchie, M. E., and Tavare, S. (2007). beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics*, **23**(16), 2183–2184.

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, **4**(8), 1184–1191.

Dutta, K. K., Zhong, Y., Liu, Y.-T., Yamada, T., Akatsuka, S., Hu, Q., Yoshihara, M., Ohara, H., *et al.* (2007). Association of microRNA-34a overexpression with proliferation is cell type-dependent. *Cancer Science*, **98**(12), 1845–1852.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, **30**(1), 207–210.

Eichner, J., Kossler, N., Wrzodek, C., Kalkuhl, A., Bach Toft, D., Ostenfeldt, N., Richard, V., and Zell, A. (2013a). A Toxicogenomic Approach for the Prediction of Murine Hepatocarcinogenesis Using Ensemble Feature Selection. *PLoS ONE*, **8**(9), e73938.

Eichner, J., Topf, F., Dräger, A., Wrzodek, C., Wanke, D., and Zell, A. (2013b). TFpredict and SABINE: sequence-based prediction of structural and functional characteristics of transcription factors. *PLoS ONE*, **8**(12), e82238.

Eichner, J., Wrzodek, C., Römer, M., Ellinger-Ziegelbauer, H., and Zell, A. (2014a). Evaluation of toxicogenomics approaches for assessing the risk of nongenotoxic carcinogenicity in rat liver. *PloS ONE*, **9**(5), e97678.

Eichner, J., Heubach, Y., Ruff, M., Kohlhof, H., Strobl, S., Mayer, B., Pawlak, M., Templin, M. F., and Zell, A. (2014b). RPPApipe: A pipeline for the analysis of reverse-phase protein array data. *Biosystems*, **122**, 19–24.

Ellinger-Ziegelbauer, H., Stuart, B., Wahle, B., Bomann, W., and Ahr, H. J. (2005). Comparison of the expression profiles induced by genotoxic and nongenotoxic carcinogens in rat liver. *Mutation Research*, **575**(1-2), 61–84.

Ellinger-Ziegelbauer, H., Gmuender, H., Bandenburg, A., and Ahr, H. J. (2008). Prediction of a carcinogenic potential of rat hepatocarcinogens using toxicogenomics analysis of short-term in vivo studies. *Mutation Research*, **637**(1-2), 23–39.

Ellinger-Ziegelbauer, H., Aubrecht, J., Kleinjans, J. C., and Ahr, H.-J. (2009). Application of toxicogenomics to study mechanisms of genotoxicity and carcinogenicity. *Toxicology Letters*, **186**(1), 36–44.

Feo, F., Frau, M., Tomasi, M. L., Brozzetti, S., and Pascale, R. M. (2009). Genetic and Epigenetic Control of Molecular Alterations in Hepatocellular Carcinoma. *Experimental Biology and Medicine*, **234**(7), 726–736.

Fielden, M. R., Brennan, R., and Gollub, J. (2007). A Gene Expression Biomarker Provides Early Prediction and Mechanistic Assessment of Hepatic Tumor Induction by Nongenotoxic Chemicals. *Toxicological Sciences*, **99**(1), 90–100.

Fielden, M. R., Nie, A., McMillian, M., Elangbam, C. S., Trela, B. A., Yang, Y., Dunn, R. T., Dragan, Y., *et al.* (2008). Interlaboratory Evaluation of Genomic Signatures for Predicting Carcinogenicity in the Rat. *Toxicological Sciences*, **103**(1), 28–34.

Firestein, G. S. and Manning, A. M. (1999). Signal transduction and transcription factors in rheumatic disease. *Arthritis & Rheumatism*, **42**(4), 609–621.

Fitzpatrick, R. B. (2008). CPDB: Carcinogenic Potency Database. *Medical Reference Services Quarterly*, **27**(3), 303–311.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., *et al.* (2014). Ensembl 2014. *Nucleic Acids Research*, **42**(D1), D749–D755.

Gaidatzis, D., van Nimwegen, E., Hausser, J., and Zavolan, M. (2007). Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, **8**, 69.

Gallagher, R. I. and Espina, V. (2014). Reverse phase protein arrays: mapping the path towards personalized medicine. *Molecular Diagnosis & Therapy*, **18**(6), 619–630.

Ganter, B., Tugendreich, S., Pearson, C. I., Ayanoglu, E., Baumhueter, S., Bostian, K. A., Brady, L., Browne, L. J., *et al.* (2005). Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *Journal of Biotechnology*, **119**(3), 219–244.

Ganter, B., Snyder, R. D., Halbert, D. N., and Lee, M. D. (2006). Toxicogenomics in drug discovery and development: mechanistic analysis of compound/class-dependent effects using the DrugMatrix ® database. *Pharmacogenomics*, **7**(7), 1025–1044.

Ganter, M., Bernard, T., Moretti, S., Stelling, J., and Pagni, M. (2013). MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks. *Bioinformatics*, **29**(6), 815–816.

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**(3), 307–315.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., *et al.* (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, **5**, R80.

Giffin, R., Pool, R., and Robinson, S. (2008). *Emerging Safety Science: Workshop Summary*. National Academies Press, Washington, DC.

Gilmore, T. D. (2006). Introduction to NF-κB: players, pathways, perspectives. *Oncogene*, **25**(51), 6680–6684.

Goecks, J., Nekrutenko, A., Taylor, J., and Galaxy Team, T. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, **11**(8), R86.

Gold, L. S., Sawyer, C. B., Magaw, R., Backman, G. M., de Veciana, M., Levinson, R., Hooper, N. K., Havender, W. R., *et al.* (1984). A carcinogenic potency database of the standardized results of animal bioassays. *Environmental Health Perspectives*, **58**, 9–319.

Goldstein, J. A., Hickman, P., and Kimbrough, R. D. (1973). Effects of purified and technical piperonyl butoxide on drug-metabolizing enzymes and ultrastructure of rat liver. *Toxicology and Applied Pharmacology*, **26**(3), 444–458.

Guo, L., Lobenhofer, E. K., Wang, C., Shippy, R., Harris, S. C., Zhang, L., Mei, N., Chen, T., *et al.* (2006). Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nature Biotechnology*, **24**(9), 1162–1169.

Gupta, M. K. and Misra, K. (2013). Modeling and simulation analysis of propylthiouracil (PTU), an anti-thyroid drug on thyroid peroxidase (TPO), thyroid stimulating hormone receptor (TSHR), and sodium iodide (NIS) symporter based on systems biology approach. *Network Modeling Analysis in Health Informatics and Bioinformatics*, **2**(1), 45–57.

Gusenleitner, D., Auerbach, S. S., Melia, T., Gómez, H. F., Sherr, D. H., and Monti, S. (2014). Genomic Models of Short-Term Exposure Accurately Predict Long-Term Chemical Carcinogenicity and Identify Putative Mechanisms of Action. *PLoS ONE*, **9**(7), e102579.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, **46**(1), 389–422.

Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y., and Aebersold, R. (2000). Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proceedings of the National Academy of Sciences*, **97**(17), 9390–9395.

Haimovitz-Friedman, A., Kolesnick, R. N., and Fuks, Z. (1997). Ceramide signaling in apoptosis. *British Medical Bulletin*, **53**(3), 539–553.

Hayashi, H., Shimamoto, K., Taniai, E., Ishii, Y., Morita, R., Suzuki, K., Shibutani, M., and Mitsumori, K. (2012). Liver tumor promoting effect of omeprazole in rats and its possible mechanism of action. *The Journal of Toxicological Sciences*, **37**(3), 491–501.

Hayes, J., Peruzzi, P. P., and Lawler, S. (2014). MicroRNAs in cancer: biomarkers, functions and therapy. *Trends in Molecular Medicine*, **20**(8), 460–469.

Hayes, J. D., Pulford, D. J., Ellis, E. M., McLeod, R., James, R. F., Seidegård, J., Mosialou, E., Jernström, B., and Neal, G. E. (1998). Regulation of rat glutathione S-transferase A5 by cancer chemopreventive agents: Mechanisms of inducible resistance to aflatoxin B1. *Chemico-Biological Interactions*, **111-112**, 51–67.

He, L. and Hannon, G. J. (2004). MicroRNAs: small RNAs with a big role in gene regulation. *Nature Reviews Genetics*, **5**(7), 522–531.

Hecht, S. S. (1999). Tobacco Smoke Carcinogens and Lung Cancer. *Journal of the National Cancer Institute*, **91**(14), 1194–1210.

Hernández, L. G., van Steeg, H., Luijten, M., and van Benthem, J. (2009). Mechanisms of non-genotoxic carcinogens and importance of a weight of evidence approach. *Mutation Research*, **682**(2-3), 94–109.

Hildebrandt, A. K., Stockel, D., Fischer, N. M., de la Garza, L., Kruger, J., Nickels, S., Rottig, M., Scharfe, C., *et al.* (2015). ballaxy: web services for structural bioinformatics. *Bioinformatics*, **31**(1), 121–122.

Hill, A. and Whitley, M. (2003). Quality Control of Expression Profiling Data. In M. E. Burczynski, editor, *An Introduction to Toxicogenomics*, chapter 3, pages 29–44. CRC Press, 1 edition.

Hiroshi Matsumoto (2011). New Short Term Prediction Method for Chemical Carcinogenicity by Hepatic Transcript Profiling Following 28-Day Toxicity Tests in Rats. *Cancer Informatics*, **10**, 259.

Hsu, S.-D., Lin, F.-M., Wu, W.-Y., Liang, C., Huang, W.-C., Chan, W.-L., Tsai, W.-T., Chen, G.-Z., *et al.* (2011). miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Research*, **39**(D1), D163–D169.

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., *et al.* (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**(4), 524–531.

Hurley, J. H., Dean, A. M., Thorsness, P. E., Koshland, D E, J., and Stroud, R. M. (1990). Regulation of isocitrate dehydrogenase by phosphorylation involves no long-range conformational change in the free enzyme. *Journal of Biological Chemistry*, **265**(7), 3599–3602.

IARC (1987). Steroidal estrogens. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*, **Suppl. 7**.

IARC (1999). Hexachloroethane. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*, **73**, 295–306.

IARC (2012). Tamoxifen. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*, **66**(100A).

IARC (2014). *World Cancer Report 2014*. World Health Organization, Lyon.

Iida, M., Anna, C. H., Holliday, W. M., Collins, J. B., Cunningham, M. L., Sills, R. C., and Devereux, T. R. (2005). Unique patterns of gene expression changes in liver after treatment of mice for 2 weeks with different known carcinogens and non-carcinogens. *Carcinogenesis*, **26**(3), 689–699.

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

Jackson, M. A., Frank Stack, H., and Waters, M. D. (1993). The genetic toxicology of putative nongenotoxic carcinogens. *Mutation Research*, **296**(3), 241–277.

Jacobs, A. (2005). Prediction of 2-Year Carcinogenicity Study Results for Pharmaceutical Products: How Are We Doing? *Toxicological Sciences*, **88**(1), 18–23.

Jiang, L., Koch, A., and Zell, A. (2016). Object Recognition and Tracking for Indoor Robots Using an RGB-D Sensor. In *Proceedings of the 13th International Conference IAS-13*, pages 859–871.

Johnson, D. E. (2012). Estimating Human Cancer Risk from Rodent Carcinogenicity Studies: The Changing Paradigm for Pharmaceuticals. *Journal of Drug Metabolism & Toxicology*, **03**(06).

Johnson, L. N. (2009a). Protein kinase inhibitors: contributions from structure to clinical compounds. *Quarterly Reviews of Biophysics*, **42**(01), 1.

Johnson, L. N. (2009b). The regulation of protein phosphorylation. *Biochemical Society Transactions*, **37**(4), 627–641.

Johnson, L. N. and Lewis, R. J. (2001). Structural Basis for Control by Phosphorylation. *Chemical Reviews*, **101**(8), 2209–2242.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., *et al.* (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**(9), 1236–1240.

Jonker, M. J., Bruning, O., van Iterson, M., Schaap, M. M., van der Hoeven, T. V., Vrieling, H., Beems, R. B., de Vries, A., *et al.* (2009). Finding transcriptomics biomarkers for in vivo identification of (non-)genotoxic carcinogens using wild-type and Xpa/p53 mutant mouse models. *Carcinogenesis*, **30**(10), 1805–12.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, **28**(1), 27–30.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, **40**(D1), D109–D114.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research*, **42**(D1), D199–D205.

Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B., Assad-Garcia, N., Glass, J. I., and Covert, M. W. (2012). A Whole-Cell Computational Model Predicts Phenotype from Genotype. *Cell*, **150**(2), 389–401.

Kasinski, A. L. and Slack, F. J. (2011). MicroRNAs en route to the clinic: progress in validating and targeting microRNAs for cancer therapy. *Nature Reviews Cancer*, **11**(12), 849–864.

Kauffmann, A., Gentleman, R., and Huber, W. (2009). arrayQualityMetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics*, **25**(3), 415–416.

Kew, M. C. (2003). Synergistic interaction between aflatoxin B1 and hepatitis B virus in hepatocarcinogenesis. *Liver International*, **23**(6), 405–409.

Khan, J., Saal, L. H., Bittner, M. L., Chen, Y., Trent, J. M., and Meltzer, P. S. (1999). Expression profiling in cancer using cDNA microarrays. *Electrophoresis*, **20**(2), 223–229.

Khan, S. R., Baghdasarian, A., Fahlman, R. P., Michail, K., and Siraki, A. G. (2014). Current status and future prospects of toxicogenomics in drug discovery. *Drug Discovery Today*, **19**(5), 562–578.

King, Z. A. (2015). BiGG Models ID Specification and Guidelines. `https://github.com/SBRG/bigg_models/wiki/`, accessed 07 January 2015.

King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O., and Lewis, N. E. (2016). BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Research*, **44**(D1), D515–D522.

Kitano, H. (2002). Computational systems biology. *Nature*, **420**(6912), 206–210.

Knight, A., Bailey, J., and Balcombe, J. (2006a). Animal carcinogenicity studies: 2. Obstacles to extrapolation of data to humans. *ATLA Alternatives to Laboratory Animals*, **34**(1), 29–38.

Knight, A., Bailey, J., and Balcombe, J. (2006b). Animal carcinogenicity studies: 3. Alternatives to the bioassay. *ATLA Alternatives to Laboratory Animals*, **34**(1), 39–48.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence*, **14**, 1137–1145.

Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y. A., Williams, E., Dylag, M., Kurbatova, N., *et al.* (2015). ArrayExpress update–simplifying data submissions. *Nucleic Acids Research*, **43**(D1), D1113–D1116.

Konig, M., Drager, A., and Holzhutter, H.-G. (2012). CySBML: a Cytoscape plugin for SBML. *Bioinformatics*, **28**(18), 2402–2403.

Kossler, N., Matheis, K. A., Ostenfeldt, N., Bach Toft, D., Dhalluin, S., Deschl, U., and Kalkuhl, A. (2015). Identification of specific mRNA signatures as fingerprints for carcinogenesis in mice induced by genotoxic and nongenotoxic hepatocarcinogens. *Toxicological Sciences*, **143**(2), 277–295.

Kozak, M. (1999). Initiation of translation in prokaryotes and eukaryotes. *Gene*, **234**(2), 187–208.

Kramer, J. A., Curtiss, S. W., Kolaja, K. L., Alden, C. L., Blomme, E. A. G., Curtiss, W. C., Davila, J. C., Jackson, C. J., and Bunch, R. T. (2004). Acute Molecular Markers of Rodent Hepatic Carcinogenesis Identified by Transcription Profiling. *Chemical Research in Toxicology*, **17**(4), 463–470.

Krause, F., Uhlendorf, J., Lubitz, T., Schulz, M., Klipp, E., and Liebermeister, W. (2010). Annotation and merging of SBML models with semanticSBML. *Bioinformatics*, **26**(3), 421–422.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, **28**(5).

Kumar, N. M. and Gilula, N. B. (1996). The Gap Junction Communication Channel. *Cell*, **84**(3), 381–388.

Lagos-Quintana, M., Rauhut, R., Meyer, J., Borkhardt, A., and Tuschl, T. (2003). New microRNAs from mouse and human. *RNA*, **9**(2), 175–179.

Lang, R. and Redmann, U. (1979). Non-mutagenicity of some sex hormones in the ames salmonella/microsome mutagenicity test. *Mutation Research*, **67**(4), 361–365.

Laskov, P. and Šrndić, N. (2011). Static detection of malicious JavaScript-bearing PDF documents. In *Proceedings of the 27th Annual Computer Security Applications Conference on - ACSAC '11*, page 373, New York, New York, USA. ACM Press.

Laulederkind, S. J. F., Hayman, G. T., Wang, S.-J., Smith, J. R., Lowry, T. F., Nigam, R., Petri, V., de Pons, J., *et al.* (2013). The Rat Genome Database 2013–data, tools and users. *Briefings in Bioinformatics*, **14**(4), 520–526.

Lee, Y. (2002). MicroRNA maturation: stepwise processing and subcellular localization. *The EMBO Journal*, **21**(17), 4663–4670.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, **11**(10), 733–739.

Lefevre, P., Tinwell, H., Galloway, S., Hill, R., Mackay, J., Elcombe, C., Foster, J., Randall, V., Callander, R., and Ashby, J. (1994). Evaluation of the Genetic Toxicity of the Peroxisome Proliferator and Carcinogen Methyl Clofenapate, Including Assays Usin Muta TM Mouse and Big BlueTM Transgenic Mice. *Human & Experimental Toxicology*, **13**(11), 764–775.

Lempiäinen, H., Müller, A., Brasa, S., Teo, S.-S., Roloff, T.-C., Morawiec, L., Zamurovic, N., Vicart, A., *et al.* (2011). Phenobarbital Mediates an Epigenetic Switch at the Constitutive Androstane Receptor (CAR) Target Gene Cyp2b10 in the Liver of B6C3F1 Mice. *PLoS ONE*, **6**(3), e18216.

Lempiainen, H., Couttet, P., Bolognani, F., Muller, A., Dubost, V., Luisier, R., del Rio-Espinola, A., Vitry, V., *et al.* (2013). Identification of Dlk1-Dio3 Imprinted Gene Cluster Noncoding RNAs as Novel Candidate Biomarkers for Liver Tumor Promotion. *Toxicological Sciences*, **131**(2), 375–386.

Lenoir, T. and Giannella, E. (2006). The emergence and diffusion of DNA microarray technology. *Journal of Biomedical Discovery and Collaboration*, **1**(1), 11.

Levandowsky, M. and Winter, D. (1971). Distance between Sets. *Nature*, **234**(5323), 34–35.

Lewis, B. P., Shih, I.-h., Jones-Rhoades, M. W., Bartel, D. P., and Burge, C. B. (2003). Prediction of Mammalian MicroRNA Targets. *Cell*, **115**(7), 787–798.

Li, C., Donizelli, M., Rodriguez, N., Dharuri, H., Endler, L., Chelliah, V., Li, L., He, E., *et al.* (2010). BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology*, **4**(92).

Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., and Cui, Q. (2014). HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Research*, **42**(D1), D1070–D1074.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News*, **2/3**.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**(12), 1739–1740.

Lijinsky, W., Reuber, M., and Blackwell, B. (1980). Liver tumors induced in rats by oral administration of the antihistaminic methapyrilene hydrochloride. *Science*, **209**(4458), 817–819.

Lim, L. P. (2003). The microRNAs of Caenorhabditis elegans. *Genes & Development*, **17**(8), 991–1008.

List, M., Block, I., Pedersen, M. L., Christiansen, H., Schmidt, S., Thomassen, M., Tan, Q., Baumbach, J., and Mollenhauer, J. (2014). Microarray R-based analysis of complex lysate experiments with MIRACLE. *Bioinformatics*, **30**(17), i631–i638.

Liu, B., Xu, J., Lan, X., Xu, R., Zhou, J., Wang, X., and Chou, K.-C. (2014). iDNA-Prot|dis: Identifying DNA-Binding Proteins by Incorporating Amino Acid Distance-Pairs and Reduced Alphabet Profile into the General Pseudo Amino Acid Composition. *PLoS ONE*, **9**(9), e106691.

Liu, X.-e., Dewaele, S., Vanhooren, V., Fan, Y.-D., Wang, L., Van Huysse, J., Zhuang, H., Contreras, R., Libert, C., and Chen, C. C. (2010). Alteration of N-glycome in diethylnitrosamine-induced hepatocellular carcinoma mice: a non-invasive monitoring tool for liver cancer. *Liver International*, **30**(8), 1221–1228.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., *et al.* (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, **14**(13), 1675–1680.

Lönnstedt, I. and Speed, T. (2001). Replicated Microarray Data. *Statistica Sinica*, **12**, 31–46.

Lopes, U. G., Erhardt, P., Yao, R., and Cooper, G. M. (1997). p53-dependent Induction of Apoptosis by Proteasome Inhibitors. *Journal of Biological Chemistry*, **272**(20), 12893–12896.

López-Romero, P. (2011). Pre-processing and differential expression analysis of Agilent microRNA arrays using the AgiMicroRna Bioconductor library. *BMC Genomics*, **12**(1), 64.

López-Romero, P., González, M. a., Callejas, S., Dopazo, A., and Irizarry, R. a. (2010). Processing of Agilent microRNA array data. *BMC Research Notes*, **3**, 18.

Lu, J. and Clark, A. G. (2012). Impact of microRNA regulation on variation in human gene expression. *Genome Research*, **22**(7), 1243–1254.

Luisier, R., Lempiainen, H., Scherbichler, N., Braeuning, A., Geissler, M., Dubost, V., Muller, A., Scheer, N., *et al.* (2014). Phenobarbital Induces Cell Cycle Transcriptional Responses in Mouse Liver Humanized for Constitutive Androstane and Pregnane X Receptors. *Toxicological Sciences*, **139**(2), 501–511.

Magee, P. N. (1969). In vivo reactions of nitroso compounds. *Annals of the New York Academy of Sciences*, **163**(2), 717–729.

Magkoufopoulou, C., Claessen, S. M. H., Tsamou, M., Jennen, D. G. J., Kleinjans, J. C. S., and van Delft, J. H. M. (2012). A transcriptomics-based in vitro assay for predicting chemical genotoxicity in vivo. *Carcinogenesis*, **33**(7), 1421–9.

Maglott, D. (2004). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, **33**(D1), D54–D58.

Mandell, D. J., Chorny, I., Groban, E. S., Wong, S. E., Levine, E., Rapp, C. S., and Jacobson, M. P. (2007). Strengths of Hydrogen Bonds Involving Phosphorylated Amino Acid Side Chains. *Journal of the American Chemical Society*, **129**(4), 820–827.

Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The Protein Kinase Complement of the Human Genome. *Science*, **298**(5600), 1912–1934.

Maragkakis, M., Reczko, M., Simossis, V. A., Alexiou, P., Papadopoulos, G. L., Dalamagas, T., Giannopoulos, G., Goumas, G., *et al.* (2009). DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Research*, **37**(W1), W273–W276.

MARCAR Consortium (2010). The Project - MARCAR. `http://www.imi-marcar.eu/project.html`, accessed 01 September 2015.

Martelli, A., Campart, G. B., Ghia, M., Allavena, A., Mereto, E., and Brambilla, G. (1996). Induction of micronuclei and initiation of enzyme-altered foci in the liver of female rats treated with cyproterone acetate, chlormadinone acetate or megestrol acetate. *Carcinogenesis*, **17**(3), 551–554.

Mastrocola, R., Aragno, M., Betteto, S., Brignardello, E., Catalano, M. G., Danni, O., and Boccuzzi, G. (2003). Pro-oxidant effect of dehydroepiandrosterone in rats is mediated by PPAR activation. *Life Sciences*, **73**(3), 289–299.

Matsumoto, H., Yakabe, Y., Saito, K., Sumida, K., Sekijima, M., Nakayama, K., Miyaura, H., Saito, F., Otsuka, M., and Shirai, T. (2009). Discrimination of carcinogens by hepatic transcript profiling in rats following 28-day administration. *Cancer Informatics*, **7**, 253–69.

Matsumoto, H., Saito, F., and Takeyoshi, M. (2014). CARCINOscreen®: New short-term prediction method for hepatocarcinogenicity of chemicals based on hepatic transcript profiling in rats. *The Journal of toxicological sciences*, **39**(5), 725–34.

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., *et al.* (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Research*, **37**(D1), D619–D622.

Matys, V. (2003). TRANSFAC(R): transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, **31**(1), 374–378.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., *et al.* (2006). TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, **34**(D1), D108–D110.

Mazur, P. K., Reynoird, N., Khatri, P., Jansen, P. W. T. C., Wilkinson, A. W., Liu, S., Barbash, O., Van Aller, G. S., *et al.* (2014). SMYD3 links lysine methylation of MAP3K2 to Ras-driven cancer. *Nature*, **510**(7504), 283–287.

McCann, J., Choi, E., Yamasaki, E., and Ames, B. N. (1975). Detection of carcinogens as mutagens in the Salmonella/microsome test: assay of 300 chemicals. *Proceedings of the National Academy of Sciences*, **72**(12), 5135–5139.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, **5**(4), 115–133.

McMillian, M., Nie, A. Y., Parker, J. B., Leone, A., Bryant, S., Kemmerer, M., Herlich, J., Liu, Y., *et al.* (2004). A gene expression signature for oxidant stress/reactive metabolites in rat liver. *Biochemical Pharmacology*, **68**(11), 2249–2261.

Melis, J. P. M., Derks, K. W. J., Pronk, T. E., Wackers, P., Schaap, M. M., Zwart, E., van Ijcken, W. F. J., Jonker, M. J., *et al.* (2014). In vivo murine hepatic microRNA and mRNA expression signatures predicting the (non-)genotoxic carcinogenic potential of chemicals. *Archives of toxicology*, **88**(4), 1023–34.

Melnick, R. L., Kohn, M. C., and Portier, C. J. (1996). Implications for risk assessment of suggested nongenotoxic mechanisms of chemical carcinogenesis. *Environmental Health Perspectives*, **104**(Suppl 1), 123–134.

Miklos, G. L. G. and Maleszka, R. (2001). Protein functions and biological contexts. *Proteomics*, **1**(2), 169–178.

Miller, E. C. and Miller, J. A. (1981). Mechanisms of chemical carcinogenesis. *Cancer*, **47**(S1), 1055–1064.

Mirkova, E. T. (1994). Activity of the rodent carcinogen 1,4-dioxane in the mouse bone marrow micronucleus assay. *Mutation Research*, **322**(2), 142–144.

Monro, A. and Mordenti, J. (1995). Expression of Exposure in Negative Carcinogenicity Studies: Dose/Body Weight, Dose/Body Surface Area, or Plasma Concentrations? *Toxicologic Pathology*, **23**(2), 187–198.

Morales-Ibanez, O., Affò, S., Rodrigo-Torres, D., Blaya, D., Millán, C., Coll, M., Perea, L., Odena, G., *et al.* (2016). Kinase analysis in alcoholic hepatitis identifies p90RSK as a potential mediator of liver fibrogenesis. *Gut*, **65**(5), 840–851.

Narang, P., Khan, S., Hemrom, A., and Lynn, A. (2014). MetaNET - a web-accessible interactive platform for biological metabolic network analysis. *BMC Systems Biology*, **8**(130).

Nie, A. Y., McMillian, M., Brandon Parker, J., Leone, A., Bryant, S., Yieh, L., Bittner, A., Nelson, J., *et al.* (2006). Predictive toxicogenomics approaches reveal underlying molecular mechanisms of nongenotoxic carcinogenicity. *Molecular Carcinogenesis*, **45**(12), 914–933.

Nishimura, D. (2001). BioCarta. *Biotech Software & Internet Report*, **2**(3), 117–120.

Nishimura, J., Dewa, Y., Okamura, T., Muguruma, M., Jin, M., Saegusa, Y., Umemura, T., and Mitsumori, K. (2008). Possible involvement of oxidative stress in fenofibrate-induced hepatocarcinogenesis in rats. *Archives of Toxicology*, **82**(9), 641–654.

Nystrom-Persson, J., Igarashi, Y., Ito, M., Morita, M., Nakatsu, N., Yamada, H., and Mizuguchi, K. (2013). Toxygates: interactive toxicity analysis on a hybrid microarray and linked data platform. *Bioinformatics*, **29**(23), 3080–3086.

Obeid, L., Linardic, C., Karolak, L., and Hannun, Y. (1993). Programmed cell death induced by ceramide. *Science*, **259**(5102), 1769–1771.

Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006). Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks. *Cell*, **127**(3), 635–648.

Olsen, P. H. and Ambros, V. (1999). The lin-4 Regulatory RNA Controls Developmental Timing in Caenorhabditis elegans by Blocking LIN-14 Protein Synthesis after the Initiation of Translation. *Developmental Biology*, **216**(2), 671–680.

Onakpoya, I. J., Heneghan, C. J., and Aronson, J. K. (2016). Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature. *BMC Medicine*, **14**(1), 10.

Paik, W. K., Paik, D. C., and Kim, S. (2007). Historical review: the field of protein methylation. *Trends in Biochemical Sciences*, **32**(3), 146–152.

Pan, J.-B., Hu, S.-C., Wang, H., Zou, Q., and Ji, Z.-L. (2012). PaGeFinder: quantitative identification of spatiotemporal pattern genes. *Bioinformatics*, **28**(11), 1544–1545.

Pan, J.-B., Hu, S.-C., Shi, D., Cai, M.-C., Li, Y.-B., Zou, Q., and Ji, Z.-L. (2013). PaGenBase: A Pattern Gene Database for the Global and Dynamic Understanding of Gene Function. *PLoS ONE*, **8**(12), e80747.

Papadopoulos, G. L., Reczko, M., Simossis, V. A., Sethupathy, P., and Hatzigeorgiou, A. G. (2009). The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Research*, **37**(D1), D155–D158.

Parman, C., Halling, C., and Gentleman, R. (2016). affyQCReport: QC Report Generation for affyBatch objects. Technical report.

Pathak, D. N. and Bodell, W. J. (1994). DNA adduct formation by tamoxifen with rat and human liver microsomal activation systems. *Carcinogenesis*, **15**(3), 529–532.

Pathak, R. K., Taj, G., Pandey, D., Arora, S., and Kumar, A. (2013). Modeling of the MAPK machinery activation in response to various abiotic and biotic stresses in plants by a system biology approach. *Bioinformation*, **9**(9), 443–449.

Patterson, T. A., Lobenhofer, E. K., Fulmer-Smentek, S. B., Collins, P. J., Chu, T.-M., Bao, W., Fang, H., Kawasaki, E. S., *et al.* (2006). Performance comparison of one-color and two-color platforms within the Microarray Quality Control (MAQC) project. *Nature Biotechnology*, **24**(9), 1140–1150.

Pawlak, M., Schick, E., Bopp, M. A., Schneider, M. J., Oroszlan, P., and Ehrat, M. (2002). Zeptosens' protein microarrays: A novel high performance microarray platform for low abundance protein analysis. *Proteomics*, **2**(4), 383.

Peraza, M. A. (2005). The Toxicology of Ligands for Peroxisome Proliferator-Activated Receptors (PPAR). *Toxicological Sciences*, **90**(2), 269–295.

Pirnia, F., Pawlak, M., Thallinger, G. G., Gierke, B., Templin, M. F., Kappeler, A., Betticher, D. C., Gloor, B., and Borner, M. M. (2009). Novel functional profiling approach combining reverse phase protein microarrays and human 3-D ex vivo tissue cultures: Expression of apoptosis-related proteins in human colon cancer. *Proteomics*, **9**(13), 3535–3548.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. `https://www.r-project.org/`.

Rahman, I. and MacNee, W. (1998). Role of transcription factors in inflammatory lung diseases. *Thorax*, **53**(7), 601–612.

Riegler, T., Nejabat, M., Eichner, J., Stiebellehner, M., Subosits, S., Bilban, M., Zell, A., Huber, W. W., Schulte-Hermann, R., and Grasl-Kraupp, B. (2015). Proinflammatory mesenchymal effects of the non-genotoxic hepatocarcinogen phenobarbital: a novel mechanism of antiapoptosis and tumor promotion. *Carcinogenesis*, **36**(12), 1521–1530.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge University Press, Cambridge.

Rivedal, E. and Witz, G. (2005). Metabolites of benzene are potent inhibitors of gap-junction intercellular communication. *Archives of Toxicology*, **79**(6), 303–311.

Roberts, R. A. and Kimber, I. (1999). Cytokines in non-genotoxic hepatocarcinogenesis. *Carcinogenesis*, **20**(8), 1397–1402.

Rogers, D. and Hahn, M. (2010). Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, **50**(5), 742–754.

Römer, M., Eichner, J., Metzger, U., Templin, M. F., Plummer, S., Ellinger-Ziegelbauer, H., and Zell, A. (2014a). Cross-platform toxicogenomics for the prediction of non-genotoxic hepatocarcinogenesis in rat. *PLoS ONE*, **9**(5), e97640.

Römer, M., Backert, L., Eichner, J., and Zell, A. (2014b). ToxDBScan: Large-Scale Similarity Screening of Toxicological Databases for Drug Candidates. *International Journal of Molecular Sciences*, **15**(10), 19037–19055.

Römer, M., Ellinger-Ziegelbauer, H., Grasl-Kraupp, B., Schwarz, M., and Zell, A. (2016a). MARCARviz: Interactive web-platform for exploratory analysis of toxicogenomics data for nongenotoxic hepatocarcinogenesis. *PeerJ Preprints*, **4**, e2393v1.

Römer, M., Eichner, J., Dräger, A., Wrzodek, C., Wrzodek, F., and Zell, A. (2016b). ZBIT Bioinformatics Toolbox: A Web-Platform for Systems Biology and Expression Data Analysis. *PLoS ONE*, **11**(2), e0149263.

Russo, A. A., Jeffrey, P. D., and Pavletich, N. P. (1996). Structural basis of cyclin-dependent kinase activation by phosphorylation. *Nature Structural Biology*, **3**(8), 696–700.

Safe, S. (1989). Polychlorinated biphenyls (PCBs): mutagenicity and carcinogenicity. *Mutation Research*, **220**(1), 31–47.

Sakai, M., Okuda, A., Hatayama, I., Sato, K., Nishi, S., and Muramatsu, M. (1989). Structure and expression of the rat c-jun messenger RNA: tissue distribution and increase during chemical hepatocarcinogenesis. *Cancer Research*, **49**(20), 5633–5637.

Salnikow, K. and Zhitkovich, A. (2008). Genetic and Epigenetic Mechanisms in Metal Carcinogenesis and Cocarcinogenesis: Nickel, Arsenic, and Chromium. *Chemical Research in Toxicology*, **21**(1), 28–44.

Schaap, M. M., Zwart, E. P., Wackers, P. F. K., Huijskens, I., van de Water, B., Breit, T. M., van Steeg, H., Jonker, M. J., and Luijten, M. (2012). Dissecting modes of action of non-genotoxic carcinogens in primary mouse hepatocytes. *Archives of toxicology*, **86**(11), 1717–27.

Schaap, M. M., Wackers, P. F. K., Zwart, E. P., Huijskens, I., Jonker, M. J., Hendriks, G., Breit, T. M., van Steeg, H., van de Water, B., and Luijten, M. (2014). A novel toxicogenomics-based approach to categorize (non-)genotoxic carcinogens. *Archives of toxicology*.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K. H. (2009). PID: the Pathway Interaction Database. *Nucleic Acids Research*, **37**(D1), D674–D679.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, **270**(5235), 467–470.

Schmitz-Spanke, S. and Rettenmeier, A. W. (2011). Protein expression profiling in chemical carcinogenesis: A proteomic-based approach. *Proteomics*, **11**(4), 644–656.

Schröder, A., Wollnik, J., Wrzodek, C., Dräger, A., Bonin, M., Burk, O., Thomas, M., Thasler, W. E., Zanger, U. M., and Zell, A. (2011). Inferring statin-induced gene regulatory relationships in primary human hepatocytes. *Bioinformatics*, **27**(18), 2473–2477.

Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., and Herzel, H. (2000). Normalization strategies for cDNA microarrays. *Nucleic Acids Research*, **28**(10), e47.

Schulte-Hermann, R., Hoffman, V., Parzefall, W., Kallenbach, M., Gerhardt, A., and Schuppler, J. (1980). Adaptive responses of rat liver to the gestagen and anti-androgen cyproterone acetate and other inducers. II. Induction of growth. *Chemico-Biological Interactions*, **31**(3), 287–300.

Shahbazian, M. D. and Grunstein, M. (2007). Functions of Site-Specific Histone Acetylation and Deacetylation. *Annual Review of Biochemistry*, **76**(1), 75–100.

Shen, S. Y., Bergmann, F., and Sauro, H. M. (2010). SBML2TikZ: supporting the SBML render extension in LaTeX. *Bioinformatics*, **26**(21), 2794–2795.

Shi, L., Tong, W., Fang, H., Scherf, U., Han, J., Puri, R. K., Frueh, F. W., Goodsaid, F. M., *et al.* (2005). Cross-platform comparability of microarray technology: Intra-platform consistency and appropriate data analysis procedures are essential. *BMC Bioinformatics*, **6**(Suppl 2), S12.

Shi, L., Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., *et al.* (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, **24**(9), 1151–1161.

Sierra-Santoyo, A., Hernández, M., Albores, A., and Cebrián, M. E. (2000). Sex-Dependent Regulation of Hepatic Cytochrome P-450 by DDT. *Toxicological Sciences*, **54**(1), 81–87.

Silva Lima, B. and Van der Laan, J. W. (2000). Mechanisms of Nongenotoxic Carcinogenesis and Assessment of the Human Hazard. *Regulatory Toxicology and Pharmacology*, **32**(2), 135–143.

Silvera, D., Formenti, S. C., and Schneider, R. J. (2010). Translational control in cancer. *Nature Reviews Cancer*, **10**(4), 254–266.

Škuta, C., BartÅŕněk, P., and Svozil, D. (2014). InCHlib âĂŞ interactive cluster heatmap for web applications. *Journal of Cheminformatics*, **6**(1), 44.

Smyth, G. K. (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**(1), 1–25.

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

Song, L., Li, D., Zeng, X., Wu, Y., Guo, L., and Zou, Q. (2014). nDNA-prot: identification of DNA-binding proteins based on unbalanced classification. *BMC Bioinformatics*, **15**(1), 298.

Sonnenburg, S., Rätsch, G., Henschel, S., Widmer, C., Behr, J., Zien, A., de Bona, F., Binder, A., Gehl, C., and Franc, V. (2010). The SHOGUN Machine Learning Toolbox. *The Journal of Machine Learning Research*, **11**, 1799–1802.

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

Stickney, J. A., Sager, S. L., Clarkson, J. R., Smith, L. A., Locey, B. J., Bock, M. J., Hartung, R., and Olp, S. F. (2003). An updated evaluation of the carcinogenic potential of 1,4-dioxane. *Regulatory Toxicology and Pharmacology*, **38**(2), 183–195.

Takashima, K., Mizukawa, Y., Morishita, K., Okuyama, M., Kasahara, T., Toritsuka, N., Miyagishima, T., Nagao, T., and Urushidani, T. (2006). Effect of the difference in vehicles on gene expression in the rat liverâĂŤanalysis of the control data in the Toxicogenomics Project Database. *Life Sciences*, **78**(24), 2787–2796.

The UniProt Consortium (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, **42**(D1), D191–D198.

Thomas, R. S., Pluta, L., Yang, L., and Halsey, T. A. (2007). Application of Genomic Biomarkers to Predict Increased Lung Tumor Incidence in 2-Year Rodent Cancer Bioassays. *Toxicological Sciences*, **97**(1), 55–64.

Thomas, R. S., Bao, W., Chu, T.-M., Bessarabova, M., Nikolskaya, T., Nikolsky, Y., Andersen, M. E., and Wolfinger, R. D. (2009). Use of Short-term Transcriptional Profiles to Assess the Long-term Cancer-Related Safety of Environmental and Industrial Chemicals. *Toxicological Sciences*, **112**(2), 311–321.

Thomas, R. S., Clewell, H. J., Allen, B. C., Wesselkamper, S. C., Wang, N. C. Y., Lambert, J. C., Hess-Wilson, J. K., Zhao, Q. J., and Andersen, M. E. (2011). Application of Transcriptional Benchmark Dose Values in Quantitative Cancer and Noncancer Risk Assessment. *Toxicological Sciences*, **120**(1), 194–205.

Thomson, J. P., Lempiäinen, H., Hackett, J. A., Nestor, C. E., Müller, A., Bolognani, F., Oakeley, E. J., Schübeler, D., *et al.* (2012). Non-genotoxic carcinogen exposure induces defined changes in the 5-hydroxymethylome. *Genome Biology*, **13**(10), R93.

Thomson, J. P., Moggs, J. G., Wolf, C. R., and Meehan, R. R. (2014). Epigenetic profiles as defined signatures of xenobiotic exposure. *Mutation Research*, **764**, 3–9.

Tong, W., Fang, H., and Mendrick, D. (2009). Toxicogenomics and Cell-based Assays for Toxicology. *Interdisciplinary Bio Central*, **1**(3), 10.1–10.5.

Tryndyak, V. P., Ross, S. A., Beland, F. A., and Pogribny, I. P. (2009). Down-regulation of the microRNAs miR-34a , miR-127 , and miR-200b in rat liver during hepatocarcinogenesis induced by a methyl-deficient diet. *Molecular Carcinogenesis*, **48**(6), 479–487.

Tucker, M. J., Adam, H. K., and Patterson, J. S. (1984). Tamoxifen. In D. R. Laurence, A. E. M. McLean, and M. Wetherall, editors, *Safety Testing of New Drugs. Laboratory Predictions and Clinical Performance*, pages 125–161. Academic Press, London.

Twyman, R. (2001). Genes in development. In *Developmental Biology*, pages 69–76. BIOS Scientific Publishers, Oxford.

Uehara, T., Hirode, M., Ono, A., Kiyosawa, N., Omura, K. K., Shimizu, T., Mizukawa, Y., Miyagishima, T., *et al.* (2008). A toxicogenomics approach for early assessment of potential non-genotoxic hepatocarcinogenicity of chemicals in rats. *Toxicology*, **250**(1), 15–26.

Uehara, T., Ono, A., Maruyama, T., Kato, I., Yamada, H., Ohno, Y., and Urushidani, T. (2010). The Japanese toxicogenomics project: application of toxicogenomics. *Molecular Nutrition & Food Research*, **54**(2), 218–227.

Uehara, T., Minowa, Y., Morikawa, Y., Kondo, C., Maruyama, T., Kato, I., Nakatsu, N., Igarashi, Y., *et al.* (2011). Prediction model of potential hepatocarcinogenicity of rat hepatocarcinogens using a large-scale toxicogenomics database. *Toxicology and Applied Pharmacology*, **255**(3), 297–306.

Unterberger, E. B., Eichner, J., Wrzodek, C., Lempiäinen, H., Luisier, R., Terranova, R., Metzger, U., Plummer, S., *et al.* (2014). Ha-ras and β-catenin oncoproteins orchestrate metabolic programs in mouse liver tumors. *International Journal of Cancer*, **135**(7), 1574–1585.

van Delft, J., van Agen, E., van Breda, S., Herwijnen, M., Staal, Y., and Kleinjans, J. (2005). Comparison of supervised clustering methods to discriminate genotoxic from non-genotoxic carcinogens by gene expression profiling. *Mutation Research*, **575**(1-2), 17–33.

Van Oosterhout, J., Van Der Laan, J., De Waal, E., Olejniczak, K., Hilgenfeld, M., Schmidt, V., and Bass, R. (1997). The Utility of Two Rodent Species in Carcinogenic Risk Assessment of Pharmaceuticals in Europe. *Regulatory Toxicology and Pharmacology*, **25**(1), 6–17.

Vapnik, V. and Lerner, A. (1963). Pattern Recognition Using Generalized Portrait Method. *Automation and Remote Control*, **24**, 774–780.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition.

Wachter, A., Bernhardt, S., Beissbarth, T., and Korf, U. (2015). Analysis of Reverse Phase Protein Array Data: From Experimental Design towards Targeted Biomarker Discovery. *Microarrays*, **4**(4), 520–539.

Wan, F. and Lenardo, M. J. (2009). Specification of DNA Binding Activity of NF- B Proteins. *Cold Spring Harbor Perspectives in Biology*, **1**(4), a000067.

Wang, H., Ach, R. A., and Curry, B. (2006). Direct and sensitive miRNA profiling from low-input total RNA. *RNA*, **13**(1), 151–159.

Ward, J. M. (2007). The Two-Year Rodent Carcinogenesis Bioassay âĂŤ Will It Survive? *Journal of Toxicologic Pathology*, **20**(1), 13–19.

Wasserman, W. W. and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, **5**(4), 276–287.

Waters, M. D., Jackson, M., and Lea, I. (2010). Characterizing and predicting carcinogenicity and mode of action using conventional and toxicogenomics methods. *Mutation Research*, **705**(3), 184–200.

Waxman, D., Ko, A., and Walsh, C. (1983). Regioselectivity and stereoselectivity of androgen hydroxylations catalyzed by cytochrome P-450 isozymes purified from phenobarbital-induced rat liver. *Journal of Biological Chemistry*, **258**(19), 11937–11947.

Weber, L. W. D., Boll, M., and Stampfl, A. (2003). Hepatotoxicity and Mechanism of Action of Haloalkanes: Carbon Tetrachloride as a Toxicological Model. *Critical Reviews in Toxicology*, **33**(2), 105–136.

Wernke, M. and Schell, J. (2004). Solvents and malignancy. *Clinics in Occupational and Environmental Medicine*, **4**(3), 513–527.

Williams, G. M. (2001). Mechanisms of chemical carcinogenesis and application to human cancer risk assessment. *Toxicology*, **166**(1-2), 3–10.

Wilson, D. N. (2013). Ribosome-targeting antibiotics and mechanisms of bacterial resistance. *Nature Reviews Microbiology*, **12**(1), 35–48.

Wittig, U., Kania, R., Golebiewski, M., Rey, M., Shi, L., Jong, L., Algaa, E., Weidemann, A., *et al.* (2012). SABIO-RK–database for biochemical reaction kinetics. *Nucleic Acids Research*, **40**(D1), D790–D796.

Wölkart, G., Schrammel, A., Dörffel, K., Haemmerle, G., Zechner, R., and Mayer, B. (2012). Cardiac dysfunction in adipose triglyceride lipase deficiency: treatment with a PPARα agonist. *British Journal of Pharmacology*, **165**(2), 380–389.

Won, M., Park, K. A., Byun, H. S., Kim, Y.-R., Choi, B. L., Hong, J. H., Park, J., Seok, J. H., *et al.* (2009). Protein kinase SGK1 enhances MEK/ERK complex formation through the phosphorylation of ERK2: Implication for the positive regulatory role of SGK1 on the ERK function during liver regeneration. *Journal of Hepatology*, **51**(1), 67–76.

Wrzodek, C., Eichner, J., Buchel, F., and Zell, A. (2013). InCroMAP: integrated analysis of cross-platform microarray and pathway data. *Bioinformatics*, **29**(4), 506–508.

Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., and Li, T. (2009). miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Research*, **37**(D1), D105–D110.

Xing, L., Wu, L., Liu, Y., Ai, N., Lu, X., and Fan, X. (2014). LTMap: a web server for assessing the potential liver toxicity by genome-wide transcriptional expression data. *Journal of Applied Toxicology*, **34**(7), 805–809.

Yang, X.-J. and Seto, E. (2008). Lysine Acetylation: Codified Crosstalk with Other Posttranslational Modifications. *Molecular Cell*, **31**(4), 449–461.

Yates, M. S., Kwak, M.-K., Egner, P. A., Groopman, J. D., Bodreddigari, S., Sutter, T. R., Baumgartner, K. J., Roebuck, B. D., *et al.* (2006). Potent Protection against Aflatoxin-Induced Tumorigenesis through Induction of Nrf2-Regulated Pathways by the Triterpenoid 1-[2-Cyano-3-,12-Dioxooleana-1,9(11)-Dien-28-Oyl]Imidazole. *Cancer Research*, **66**(4), 2488–2494.

Yildirimman, R., Brolen, G., Vilardell, M., Eriksson, G., Synnergren, J., Gmuender, H., Kamburov, A., Ingelman-Sundberg, M., *et al.* (2011). Human Embryonic Stem Cell Derived Hepatocyte-Like Cells as a Tool for In Vitro Hazard Assessment of Chemical Carcinogenicity. *Toxicological Sciences*, **124**(2), 278–290.

Yokoi, T. and Nakajima, M. (2011). Toxicological Implications of Modulation of Gene Expression by MicroRNAs. *Toxicological Sciences*, **123**(1), 1–14.

Yona, G., Dirks, W., Rahman, S., and Lin, D. M. (2006). Effective similarity measures for expression profiles. *Bioinformatics*, **22**(13), 1616–1622.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLoS ONE*, **9**(1), e78644.