Transcriptional Responses to Radiation Exposure Facilitate the Discovery of Biomarkers Functioning as Radiation Biodosimeters

Dissertation

zur

Erlangung des akademischen Grades

Doktor-Ingenieur (Dr.-Ing.)

Promotionsgebiet Bioinformatik

Fakultät für Informatik und Elektrotechnik
Universität Rostock

Universität
Rostock

gegründet 1419

Traditio et Innovatio

vorgelegt von

Sonja Strunz
(geb. Boldt)

geboren am 12. Juni 1981 in Bad Homburg v.d.H.

Rostock, 15. November 2013

urn:nbn:de:gbv:28-diss2014-0136-2

Gutachter:	Prof. Dr. Olaf Wolkenhauer, Universität Rostock
	Dr. Ralf Kriehuber, Forschungszentrum Jülich
	Prof. Dr. Mario Stanke, Universität Greifswald

Tag der Verteidigung: 13.5.2014

Wege entstehen dadurch, dass man sie geht

(Franz Kafka)

Danksagung

Ich danke Prof. Olaf Wolkenhauer, dass er mich in seine Gruppe aufgenommen hat und mir die Gelegenheit gab, unter seiner Obhut zu promovieren. Danke, dass Du in den letzten Jahren stets an mich und meine Arbeit geglaubt hast und mich gelehrt hast, Herausforderungen anzunehmen.

Diese Arbeit wäre ohne die kontinuierliche und enge Zusammenarbeit mit Dr. Ralf Kriehuber in dieser Form nicht möglich gewesen. Danke, dass Du mich bei unseren Treffen immer freundlich empfangen und fachlich unterstützt hast.

Außerdem danke ich Prof. Mario Stanke, dass er sich bereit erklärt hat, diese Arbeit zu begutachten.

In den letzten Jahren sind mir viele Menschen sehr ans Herz gewachsen, die mein Leben in vielerlei Hinsicht bereichern. Ihnen ist es zu verdanken, dass ich so manch eine Hürde bewältigt habe und meine Promotion zu einer prägenden und erfüllenden Zeit wurde.

Katja, unsere Zeit der Zusammenarbeit möchte ich nicht missen. Es ist schön, wenn aus gemeinsamer Arbeit Freundschaft wird. Ulf, ich hätte keinen besseren Bürokollegen haben können. Änne, mit Dir zusammen fing meine wissenschaftliche Arbeit in Rostock an. Unsere darauffolgenden gemeinsamen Mittagessen waren nicht nur eine wunderbare Abwechslung, sondern oft auch Motivation und Bereicherung zugleich. Stefan, egal bei welchem Problem, ich konnte immer auf Dich zählen. Florian, Deine herzliche und fröhliche Art hat immer dazu beigetragen, dass ich gerne ins Büro gekommen bin. Yvonne, vielleicht ist zusammenge*schweiß*t ein passendes Wort für uns. Was in der KTV anfing, hat auch in Hamburg nicht aufgehört. Man muss nicht benachbart sein, um füreinander da zu sein. Tina, Du hast mir oft geholfen, Dinge aus einer anderen Perspektive zu betrachten. Aus unseren Gesprächen habe ich viel Kraft geschöpft und auch gelernt, mal loszulassen. Martina, Du warst immer und in allen Lebenslagen für mich da. Jemanden an seiner Seite zu haben, der ähnlich fühlt und denkt, ist ein besonderes Glück.

Außerdem danke ich Florian, Katja, Änne, Yvonne und Nina für die Korrektur meiner Arbeit.

Letztendlich möchte ich meiner Familie danken, die mit ihrer Liebe und ihrem Vertrauen den Grundstein für meinen Weg gelegt haben. Matthias, ohne Deine Rückendeckung und Dein Vertrauen wäre ich nicht so weit gekommen. Du hast die Gabe, mich auch in schwierigen Situationen zum Lachen zu bringen und zeigst mir immer wieder wie schön das Leben ist.

Diese Arbeit wurde von dem Bundesministerium für Bildung und Forschung (BMBF) im Rahmen des Verbundprojektes "Ein systembiologischer Ansatz für die Biodosimetrie und die Analyse individueller Radiosensitivität" (02NUK005) finanziert.

Abstract

The use of ionizing radiation is a double-edged sword. On the one hand, ionizing radiation causes a broad scope of adverse human health effects, ranging from long-term effects like an increased cancer risk to short-term effects like radiation sickness. On the other hand, radiation offers many benefits to our society. Especially, the medical use of ionizing radiation for disease diagnosis and treatment, such as radiological imaging or radiotherapy, is of vital importance and is constantly increasing. Since accidental and occupational exposures have become more frequent over the last decades, the development of new methods for a retrospective quantification of the radiation dose of exposed individuals is of current widespread interest.

The primary goal of this thesis is the identification of gene expression-based signatures allowing a retrospective estimation of radiation doses after a radiation accident. To this end, I developed and implemented a bioinformatics-driven framework for biomarker discovery and radiation dose prediction. In light of recent concerns about the non-reproducibility of putative biomarkers, the algorithmic design of my computational framework intends to support the identification of gene signatures having a high stability with respect to data variations.

Ionizing radiation evokes an elaborate cellular DNA damage response including a complex transcriptional regulation. In this thesis, I first analyzed gene expression alterations in human peripheral blood lymphocytes after *ex vivo* γ-irradiation and characterized functional processes and pathways affected by low, medium and high dose exposure. My statistical and functional DNA-microarray analysis shows that (i) both the time after exposure and the radiation dose substantially influence the transcription and that (ii) even low dose exposure leads to well-defined physiological responses. By applying my computational framework to our DNA-microarray data, I successfully identified two gene signatures with which low and medium to high radiation doses can be accurately estimated respectively.

In conclusion, the results of the present work enhance our understanding of the transcriptional response induced by ionizing radiation. Furthermore, this study confirms the idea that gene expression profiles are a valuable tool for estimating even low radiation doses in a rapid and reliable manner. The results may provide the basis for a refined biodosimetry platform which can be utilized after radiation accidents to guide medical treatment in the future.

Zusammenfassung

Die Verwendung ionisierender Strahlung ist ein zweischneidiges Schwert. Einerseits besitzt ionisierende Strahlung eine umfangreiche gesundheitsschädigende Wirkung, die von Langzeitfolgen, wie erhöhtem Krebsrisiko, bis zu Kurzzeitauswirkungen, wie der Strahlenkrankheit, reicht. Andererseits bietet Strahlung großen gesellschaftlichen Nutzen. Insbesondere ihre medizinische Anwendung in der Krankheitsdiagnostik und -behandlung, wie der radiologischen Bildgebung oder Strahlentherapie, ist von zunehmender Bedeutung. Aufgrund vermehrter unfall- oder berufsbedingter Expositionen ist die Entwicklung neuer Methoden für eine retrospektive Quantifizierung der Strahlendosis exponierter Personen von großem Interesse.

Primäres Ziel dieser Dissertation ist die Identifikation von expressionsbasierten Gensignaturen für eine rückwirkende Abschätzung der individuellen Strahlendosis nach Strahlenunfällen. Zu diesem Zweck entwickelte und implementierte ich ein Framework, dessen algorithmisches Design aufgrund aktueller Bedenken über die Nichtreproduzierbarkeit potentieller Biomarker die Identifikation von Gensignaturen zur Strahlendosisvorhersage mit einer hohen Stabilität gegenüber Datenabweichungen unterstützt.

Ionisierende Strahlung ruft eine umfangreiche zelluläre DNA Schadensantwort hervor, einschließlich einer komplexen transkriptionellen Regulation. Folglich analysierte ich zunächst die Genexpressionsänderungen in humanen peripheren Lymphozyten nach *ex vivo* γ-Bestrahlung und charakterisierte die durch geringe, mittlere und hohe Strahlendosen induzierten funktionalen Prozesse und Signalwege. Meine statistische und funktionale DNA-Microarray Analyse zeigt, (i) dass sowohl die Zeit nach Exposition, als auch die Strahlendosis einen maßgeblichen Einfluss auf die Transkription ausüben und (ii) dass sogar Niedrigdosisbestrahlung zu einer klar definierten physiologischen Antwort führt. Mit Hilfe meines Frameworks identifizierte ich zwei Gensignaturen, anhand derer geringe bzw. mittlere bis hohe Strahlendosen präzise vorhergesagt werden können.

Die hier vorliegende Arbeit trägt zu einem besseren Verständnis der durch ionisierende Strahlung induzierten transkriptionellen Antwort bei. Zudem bestätigt sie die Idee, dass Genexpressionsprofile ein wertvolles Instrument darstellen, um sogar geringe Strahlendosen, schnell und zuverlässig abschätzen zu können. Letztendlich, legt sie den Grundstein für ein verfeinertes biodosimetrisches Assay, das im Falle eines Strahlenunfalls eingesetzt werden kann, um geeignete medizinische Behandlungen einzuleiten.

Contents

List of Figures

List of Tables

List of Abbreviations

$\triangle\triangle C_t$	delta delta cycle threshold
r^2	Coefficient of Determination
ANOVA	Analysis of Variance
C_t	threshold cycle value
cDNA	complementary DNA
cRNA	complementary RNA
CV	Coefficient of Variation
DAVID	Database for Annotation, Visualization and Integrated Discovery
DNA	deoxyribonucleic acid
DSB	double-strand break
FDR	false discovery rate
GAPDH	glyceraldehyde-3-phosphate dehydrogenase 1
GO	Gene Ontology
Gy	gray
h	hour(s)
KEGG	Kyoto Encyclopedia of Genes and Genomes
KNN	*k*-nearest neighbour
LET	linear energy transfer
mGy	milligray
min	minute(s)
ml	millilitre
mRNA	messenger RNA
mSv	millisievert
PBLs	peripheral blood lymphocytes
qRT-PCR	quantitative real-time polymerase chain reaction
RNA	ribonucleic acid
RNA-Seq	RNA sequencing
SAGE	serial analysis of gene expression
Sv	sievert
Tukey's HSD	Tukey's Honestly Significant Difference

Chapter 1

Introduction

Synopsis

The introduction of this interdisciplinary dissertation sets the biological and bioinformatics context of my work. The application of computational approaches to high-throughput experimental data provides a great opportunity to discover new knowledge and to gain a better understanding of the biological system being studied. I here frame the problem of biomarker discovery from high-throughput gene expression data. To this end, I present a typical workflow for the identification of biomarker signatures and summarize common approaches for each step. Furthermore, I show that this topic has the potential to support radiation research and can help to overcome limitations of current techniques for radiation biodosimetry. Additionally, I outline the scientific motivation for my research and point out the major objectives of this dissertation.

1.1 Motivation and objectives

Ionizing radiation induces many types of deoxyribonucleic acid (DNA)-lesions, which trigger a highly interwoven network of intracellular and intercellular regulatory mechanisms. The cellular response to DNA damage orchestrates major processes, such as apoptosis, cell cycle progression and DNA repair, and is accompanied by a complex transcriptional regulation. My motivation to investigate these radiation-induced transcriptional changes by gene expression profiling is twofold:

1. Measuring the gene expression accomplished by a statistical and bioinformatics-driven analysis helps to understand and characterize the functional processes and pathways involved in the DNA damage response. The molecular mechanisms triggered by low radiation doses in particular are still unclear and require further investigations.

2. Gene-based biomarkers are a promising tool for the retrospective estimation of radiation doses, which allow for a fast, minimally invasive and automated screening of a large number of exposed people. Whereas the ability to discriminate medium to high radiation doses can guide medical decision making after a large-scale radiation accident, the discrimination of low radiation doses on the basis of gene expression changes is not essential for acute medical decision making but may support the assessment of associated long-term health risks of exposed individuals. This recent branch of radiation biodosimetry is an ongoing topic of research, and problems such as the concern about the poor reproducibility of gene-based biomarkers have to be resolved. Whether gene expression dosimeters are an appropriate tool for discriminating low radiation doses is open to analysis and awaits further clarification.

In this thesis, I statistically investigated the transcriptional response in human peripheral blood lymphocytes (PBLs) to low, medium, and high dose exposure and functionally characterized the time-dependent and dose-dependent gene expression changes. This comprehensive and systematic DNA-microarray analysis is the first step towards an identification of gene expression biomarkers functioning as radiation biodosimeters. To achieve this goal, I developed and implemented a computational framework whose algorithmic design intends to support the discovery of stable biomarker genes. By applying this framework to our DNA-microarray data of irradiated human PBLs, I address the question of whether gene expression-based biomarkers are applicable for a retrospective estimation of radiation doses, and identify candidate biomarker signatures for the discrimination of low and medium to high radiation doses respectively. Ultimately, my work intends to enhance our understanding of the transcriptional response induced by ionizing radiation, especially after low dose exposure, and may provide the basis for a refined biodosimetry platform based on gene expression alterations to estimate radiation doses in radiation accidents in the future.

1.2 Microarray-based biomarker discovery

An major goal of biomedical research is the development of new approaches for diagnosis, treatment and prevention of diseases. In this context, researchers often seek biological parameters, so called biomarkers, which are indicative for specific health or disease characteristics (Vasan, 2006). As defined by the Biomarkers Definitions Working Group, a biomarker is a "characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" (Atkinson et al., 2001). The applications of such indicators are manifold: they can be utilized (a) to recognize an overt disease, (b) to screen a subclinical disease, (c) to predict the course of a disease including the patients response to therapeutic interventions, (d) to estimate the risk of developing a disease, or (d) to categorize a disease severity (Vasan, 2006). The discovery of prognostic or diagnostic biomarkers are thus a major step towards personalized medicine.
Whereas many of the already well-established, clinically relevant biomarkers are single molecular species or characteristics, modern molecular biology provides data sources for predictive signatures combining multiple molecular constituents. The discovery of molecular signatures from high-throughput "Omics" data is an active research topic in the field of bioinformatics (McDermott et al., 2013). A typical workflow for high-throughput data analysis and biomarker discovery, at which biologists and bioinformaticians are equally involved, is illustrated in Figure 1.1 on the following page. Ideally, the experimentally validated results obtained within this process enhance the understanding of the biological system under study and motivate the design of new experiments. The depicted procedure is basically applicable to many types of "Omics" data, even though single steps may vary in their implemented approaches. In what follows, each step is discussed in more detail, whereby the description is mainly dedicated to the analysis of gene expression data. Therefore, the next two sections offer a survey on the basics of gene expression profiling and the fundamentals of microarray data analysis for biomarker discovery.

1.2.1 Gene expression profiling

Gene expression is the multi-staged biological process of synthesising proteins from genes. In the first stage, the transcription, single-stranded transcripts of genes and other types of functional ribonucleic acid (RNA) molecules are produced. In Eukaryotic organisms, these transcripts of genes are called precursor messenger RNA (mRNA), which undergo post-transcriptional modifications before they are translocated from the

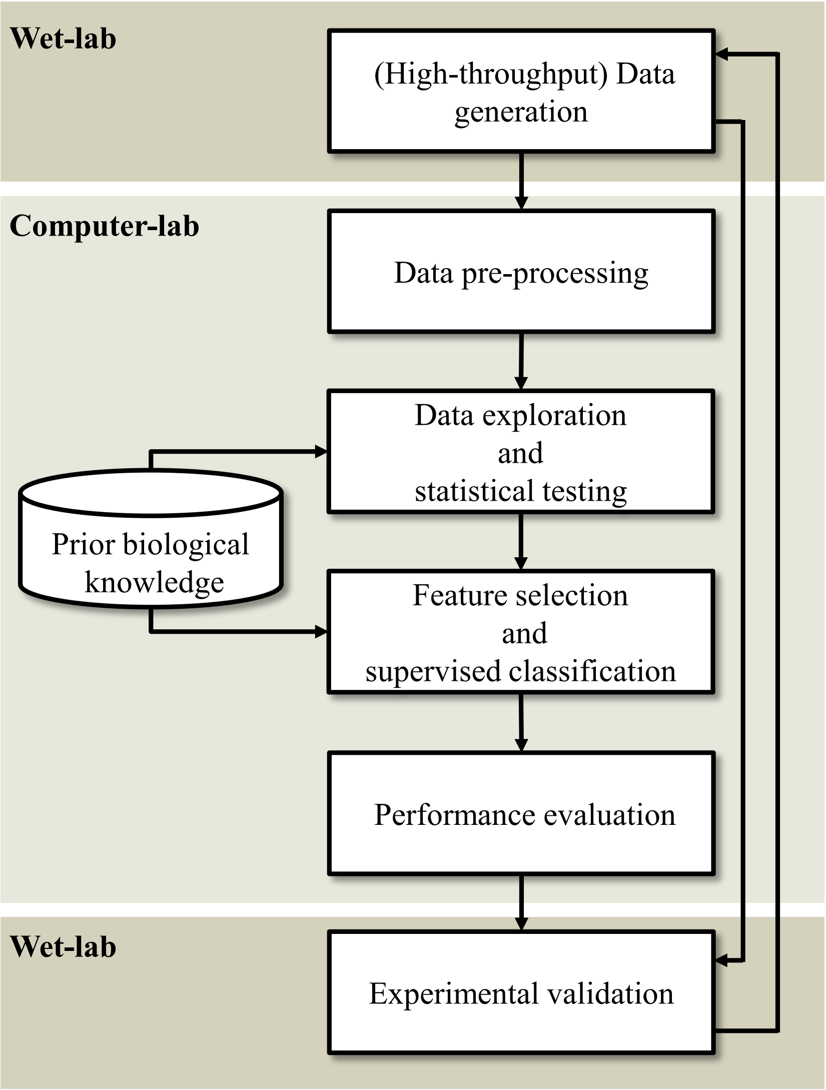

Wet-lab
(High-throughput) Data generation
Computer-lab
Data pre-processing
Data exploration and statistical testing
Prior biological knowledge
Feature selection and supervised classification
Performance evaluation
Wet-lab
Experimental validation

Figure 1.1: Generic workflow for high-throughput data analysis and biomarker discovery.

nucleus to the cytoplasm as mature mRNA molecules. The second stage of gene expression is then carried out at the ribosomes, where the mRNA molecules are translated into their corresponding amino acid sequence, the proteins. The complete set of RNA transcripts in a cell, including their quantity at a certain time point, is defined as the transcriptome (Wang et al., 2009b). The composition of the transcriptome varies for different cell types and is highly dependent on internal and external conditions of the cell (Velculescu et al., 1997).

The major aims of transcriptomics, which is the study of the cells transcriptome, are to characterize the constituents of the cell, to determine the cell- and condition-specific dynamics of transcriptional activity and to investigate regulating mechanisms of mRNA production (Jayaraman and Hahn, 2009; Wang et al., 2009b).

As recently as two decades ago, the transcriptome research was limited due to experimental techniques which could only quantify the expression levels of a small number of genes, each of them separately measured. Nowadays, various high-throughput experimental techniques have been developed, which simultaneously monitor the expression level of thousands of genes. For gene expression profiling, either hybridization-based approaches, such as microarrays, or sequence-based approaches (Holt and Jones, 2008), like serial analysis of gene expression (SAGE) (Harbers and Carninci, 2005) or the most recently developed RNA sequencing (RNA-Seq) technology, exist.

Up until now, DNA-microarrays have been extensively utilized for gene expression profiling in diverse biological contexts. The basic principle is that an ordered set of DNA fragments is positioned on a solid surface. The DNA fragments serve as probes for binding specific sequences of nucleic acids (*i.e.* targets) corresponding to particular genes of the genome. Dependent on the type of microarray, probes of different length are used for target hybridization. Common types are complementary DNA (cDNA)-microarrays, which utilize probes of hundreds or thousands of base pairs, or oligonucleotide arrays, which use shorter probes with a length of approximately 50 base pairs. The hybridized targets are labeled with a fluorescent tag and after washing, the array is scanned in order to measure the signal emitted by the labeling dye at each sequence-specific location (Jaluria et al., 2007). Assuming that the emitted signal is directly proportional to the amount of mRNA present in the sample under study, microarrays do not derive absolute levels of expression but provide a quantification which can be useful to compare different samples measuring gene expression levels under different conditions.

The advent of high-throughput technologies supported a shift from the classical reductionist approach of studying individual genes to a systems level approach, where the entire system of gene expression is considered. Although a more global view on the transcriptional states of a cell is an advantage, the wealth of expression data produced by gene expression profiling makes new demands and poses challenges for storing, retrieving and analyzing the data.

1.2.2 Computational concepts

After high-throughput data generation, there are four main steps for computational data analysis and biomarker discovery which have to be passed: (1) **data pre-processing**, (2) **data exploration and statistical testing**, (3) **feature selection and supervised classification**, and (4) **performance evaluation** (see Figure 1.1 on the preceding page).

Data pre-processing: First, the experimental raw data have to be processed. Typical issues of data pre-processing include quality assessment, removal of systematic sources of variation, scaling of raw data and the detection of outliers. For microarray experiments, the pre-processing normally consists of (a) background subtraction, which is based on the assumption that the measured signal intensity is composed of the fluorescence of the spot and some background noise, (b) scaling of signal intensities by log-transformation, and (c) data normalization. Whereas the need for background correction has been controversially discussed (Zahurak et al., 2007), scaling and normalization are inherent parts of each analysis. Microarray experiments are subject to multiple sources of technical variations, including differences in mRNA preparation, cDNA labeling or hybridization efficiency, which can considerably limit the biological interpretability of the data (Quackenbush, 2002). Normalization is a mean to adjust for such effects of technical errors within and between different arrays. A variety of different techniques for normalization has been developed, but the choice for the most appropriate method is dependent on both the context of the performed study and the array technology used for gene expression profiling (Smyth and Speed, 2003).

Data exploration and statistical testing: After pre-processing the raw data, the first step of high-throughput data analysis is often to investigate the underlying structure of the given data in an explorative way by cluster analysis. The goal of cluster analysis is to divide the measured data into groups (*i.e.* clusters) of similar data points, whereby the data points within one group are more similar to each other than data points of distinct groups. For a microarray experiment, in which genes are measured under different conditions, either genes with similar expression patterns across various conditions or samples with similar expression patterns across the measured transcriptome are identified. A review of existing clustering methods for gene expression data is given by Jain (2010).

The identification of differentially expressed genes, *e.g.* genes which show significant changes in gene expression between different conditions, is one of the basic goals of microarray experiments. To this end, statistical testing has to be carried out and often the main difficulty lies in the choice of the appropriate test statistic, which is most of all dependent on the experimental design (*e.g.* number of conditions and influencing factors, repeated measurements) and the nature of the underlying data. A detailed discussion of this issue is beyond the scope of this introduction and interested readers are referred to the review of Cui and Churchill (2003).

When testing for differentially expressed genes, two types of errors can occur: either a gene is declared as differentially expressed when it is not (*i.e.* Type I error) or a truly differentially expressed gene is not identified as such (*i.e.* Type II error). Since each statistical test has a specified Type I error probability, the chance of committing some Type I errors increases with the number of hypothesis tested (Dudoit et al., 2003). For this reason, in the context of microarray analysis, a considerable number of genes may be identified as differentially expressed simply by chance. For a scenario of a large number of simultaneously tested hypothesis, the Type I error rate can be controlled by applying multiple testing procedures. Dudoit et al. (2003) discusses different approaches for multiple hypothesis testing in the context of microarray experiments and compares the procedures on microarray and simulated datasets.

The discovery of genes with similar expression patterns across various conditions by clustering, or the identification of differentially expressed genes by statistical testing, is just an intermediate step towards the in-depth understanding of biological systems under study. The functional interpretation of the results is indispensable. The systematized knowledge about gene and protein function provided by the Gene Ontology (GO) consortium (Ashburner et al., 2000) or the information of biological pathway databases, like Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2012), both utilized in the present thesis, are just two of a myriad of available resources to derive biologically meaningful results from high-throughput data experiments.

Feature selection and supervised classification: Given a high-dimensional dataset representing two or more biological conditions (such as healthy and diseased), the goal of biomarker discovery is to find a subset of all measured features, with which the conditions under study, also referred to as classes or class labels, can be accurately predicted. Dependent on the underlying data, the features can be, for example, proteins, metabolites or, if gene expression profiling was performed in the first step, genes. If more than one feature is selected for class prediction, one often speaks of a molecular signature functioning as a biomarker. For biomarker discovery, feature selection methods in combination with supervised classification algorithms are commonly used. The basic principle behind this is that with a feature selection method a subset of features is extracted from data for which the classes are known. The measured values of these features (*e.g.* gene expression levels) together with their class labels (*i.e.* the condition under which they were measured) serve as an input for the supervised classification algorithm in order to train a classifier. Classifier training is the process of learning a set of rules or a mathematical model, which can then be used to predict the class labels of new observations. The prediction of class labels of new observations, by which the features but not the biological conditions under which they were measured are known, is called classifier prediction.

The extraction of feature subsets for classification has several advantages: it reduces the dimensionality of high-throughput data, limits the risk of overfitting during classification (see next section for explanation), supports a computationally faster classifier training and may help to gain a deeper insight into the underlying processes that generated the data (Saeys et al., 2007). The existing feature selection methods can be categorized into three main approaches, namely the filter, wrapper and embedded methods (John et al., 1994). Filter techniques score the relevance of individual features or feature subsets without incorporating any classification scheme. Typical scoring procedures include statistical testing or the analysis of relationships (*e.g.* correlations) between feature measurements and class labels. The features are ranked according to their relevance scores and a certain set of best-ranked features are selected for subsequent classifier training. Wrapper techniques, on the other hand, traverse the space of possible feature subsets and evaluate their prediction performance by applying a pre-defined supervised machine learning algorithm for classifier training. The space of possible feature subsets is often pruned by heuristic search algorithms and the subset associated with the highest performance is then selected. The third category of methods, the embedded techniques, directly integrates the search and selection of feature subsets into the process of classifier training, which means that feature selection and classifier training cannot be separated from each other. For all categories many approaches have been proposed, differing in their complexity and thereby in their interpretability. Sometimes, additional biological information is utilized to support the process of feature selection. The integration of prior biological knowledge from external repositories, such as pathway information, aims at a reduction of the dimensionality by only selecting features which are known to be relevant in the given biological context. Once a subset of features is identified, their measurements and class labels serve as an input for classifier training by applying supervised classification algorithms. Supervised classification algorithms can be statistical methods like discriminant analysis or machine learning techniques. For the latter, numerous approaches have been proposed in the last decades, which range from simple approaches like the instance-based learning methods to more complex approaches like support vector machines and random forests. Since instance-based learning, with the k-nearest neighbour (KNN) approach as its main representative, is an important component of the present work, it is explained in more detail in Section 4.1 on page 44. For a more detailed description of other machine learning algorithms, like support vector machines, decision trees, neuronal networks or Bayesian methods, including a discussion on their individual advantages and shortcomings, the reader is referred to the reviews of Kotsiantis (2007) and Larrañaga et al. (2006). The choice of the most appropriate learning algorithm is a critical step and highly dependent on the underlying data and context. Systematic evaluation studies comparing different supervised learning approaches applied to different datasets may support this decision (Dudoit et al., 2002; Lee et al., 2005).

Performance evaluation: The fourth step of the here described biomarker discovery pipeline is, as illustrated in Figure 1.1 on page 4, the evaluation of the classifier performance. The performance is a measure to describe how well a classifier discriminates the classes, *i.e.* predicts the class labels of new observations. A key concept of the performance evaluation of supervised classifications is that the performance has to be

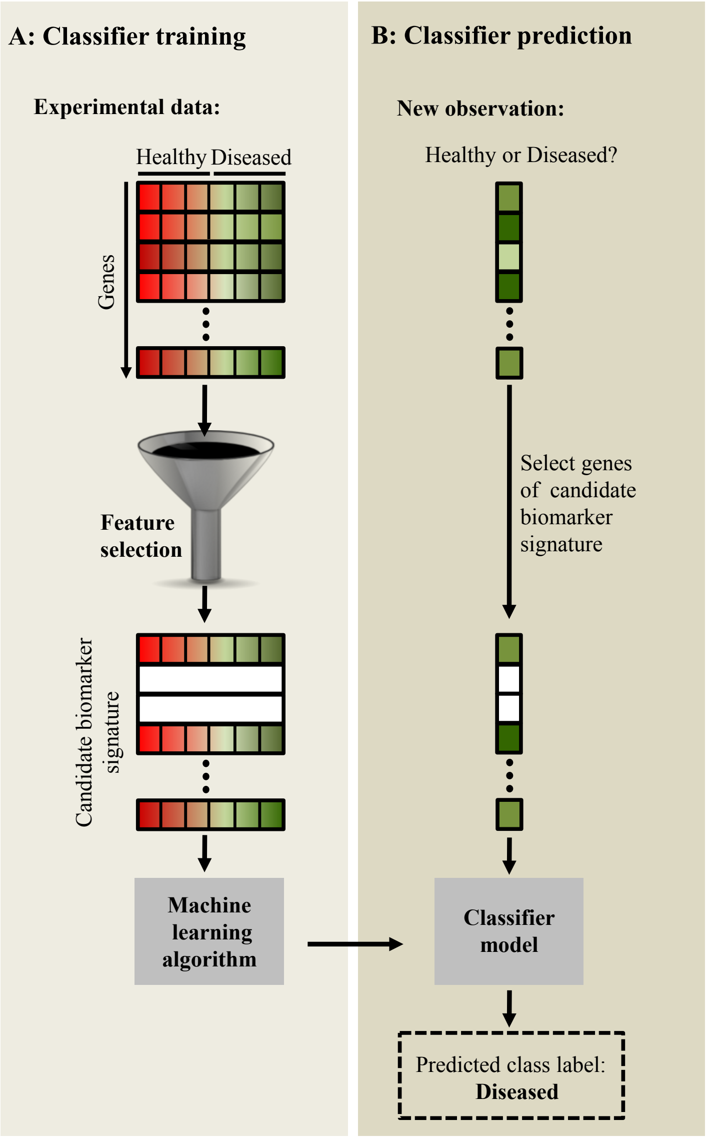


Figure 1.2: Scheme of a supervised classification procedure of using biomarker signatures for outcome prediction. A: A subset of features is selected from the given experimental data by applying a feature selection method. This subset represents the candidate biomarker signature. In the here presented example, a gene expression signature is selected from microarray samples of patients which are either healthy or diseased. The measured values of the signature (*e.g.* gene expression levels) together with their class labels (*i.e.* the condition under which they were measured like healthy and diseased) serve as an input for the supervised classification algorithm in order to train a classifier. **B:** The trained classifier model can be utilized to predict the class labels of new observations. This process is called classifier prediction. Here, the new observation is a microarray sample of a patient with unknown health status. From this all features of the candidate biomarker signature are selected. The classifier model predicts based on the measurements of the selected features the class label of the new observation (*e.g. diseased*).

evaluated with data which were not used for feature selection and classifier training before. Only when using a disjoint dataset a reliable conclusion can be drawn that the learned characteristics of the data in the training phase can be generalized to new observations. The use of distinct datasets for classifier training and performance evaluation (*i.e.* classifier testing) is often a concern in practical applications because the amount of experimental data is usually limited. Especially in high-throughput experiments, the number of measured samples, *i.e.* conditions, is far smaller than the number of measured features. Such datasets are prone to a phenomena called overfitting, meaning that the classifier appears to have a good performance on the training data but shows only limited capabilities of generalization and is therefore unable to accurately predict new observations. A common strategy to deal with the limited amount of data and with the problem of overfitting is to perform cross-validation procedures (Kohavi, 1995). Here, a classifier is trained on a subset of data (called training set) and tested on the remainder (called test set). Repeating this systematically by using different partitioning of the data, cross-validation has the potential of employing the entire training set for testing, albeit not at once, and simultaneously creating the largest possible test set for a fixed training set (Rao et al., 2008). By aggregating the classifier performances obtained for each pair of training and test sets, an overall performance for the classifier can be obtained. Depending on the way the data are partitioned into training and test sets, various types of cross-validation procedures can be distinguished. Some of the most common types are hold-out cross-validation, *k*-fold cross-validation or leave-one-out cross-validation. This splitting approach for performance evaluation is called internal validation.

After a potential biomarker signature (*i.e.* feature subset) is extracted and its ability for class prediction was evaluated by an internal validation strategy, it is recommended to employ an additional external validation. An external validation, comprising of a classifier prediction with a distinct dataset of new biological samples, supports drawing meaningful conclusions on the capability of generalization of the identified biomarker signature.

In the next step, the potential biomarker signature should be experimentally verified. For example, if a potential gene signature was identified based on microarray gene expression data, it is common practice to additionally confirm the gene expression alterations with quantitative real-time polymerase chain reaction (qRT-PCR) using an independent, newly generated dataset.

1.3 Radiation biodosimetry

Radiation is part of our daily life. Due to radioactive material in the Earth's crust and atmosphere as well as radioactive substances within our body, we are constantly exposed to radiation. Most of the radiation we receive in our life originates from these natural sources. Smaller but constantly increasing is the exposure to radiation from man-made sources. Especially in medicine, ionizing radiation has become to an essential tool for therapy and diagnosis, but also numerous industrial and agricultural applications considerably profit from its use. Since many of these man-made sources have become an integral part of our world, the risk of radiation accidents and potential health hazards has increased. This development raises not only the question of the underlying risks and health effects of radiation exposure but also substantiates the need to protect human beings from its harmful effects.

People working in an actual or potential radiation environment are supposed to wear small radiation detection devices. These personal physical dosimeters accurately monitor the external radiation dose and after an occupational exposure they provide valuable information for medical management. However, in case of a radiation accident, where a large number of persons without physical dosimeters are affected, the information of personnel monitoring measurements are either incomplete or completely lacking. In this case, methods for retrospective physical and biological dosimetry are a supplementary or even an alternative means to estimate the radiation dose already received by an individual. Hereby, biological dosimetry (*i.e.* biodosimetry) refers to the measurement of biological markers that can be quantitatively related to the magnitude of the radiation dose (Simon et al., 2010). Accurately estimating the received radiation dose of accidently exposed people can assist prompt medical decision making and might help to assess the risk of long term consequences from radiation exposure.

There is a lasting need to improve already existing techniques of radiation biodosimetry as well as to stimulate research into the development of new techniques. In the following, after briefly reviewing the fundamentals of ionizing radiation, already established methods for dose quantification are presented. Furthermore, a recent biodosimetric branch, namely the development of gene expression-based dosimeters, is introduced.

1.3.1 Ionizing radiation and its health effects

Ionizing radiation is composed of particles which have enough energy for producing charged ions when passing through matter by detaching electrons from atoms or molecules. One distinguishes between two types of ionizing radiation: the first is called particulate radiation and involves charged or uncharged fast-moving particles that have both energy and mass. Particulate radiation is primarily produced by disintegration of unstable atoms. α-radiation (*i.e.* the emission of alpha-particles consisting of two protons and two neutrons) and β-radiation (*i.e.* the emission of beta-particles consisting of either an electron or a positron) belong to this category of radiation. Whereas alpha-particles have a very short range in matter due to their large mass and can be easily shielded with paper-thin materials, beta-particles have a greater capability of penetration and shielding materials typically include aluminum (Woodside, 1997). The second type of ionizing radiation is electromagnetic radiation (photons), with energy but no mass. γ-rays, and X-rays belong to this category (Rana et al., 2010). Gamma-ray photons have no mass or charge and can penetrate tissues easily (Woodside, 1997).

The amount of ionizing radiation is defined as the energy absorbed per unit of mass. The absorbed dose is expressed in units of gray (Gy), where 1 Gy equals 1 joule of energy absorbed per kilogram of matter (Cameron, 1991). The biological damage caused by a particle depends not only on the received radiation dose but also on its linear energy transfer (LET), which is defined as the rate of energy loss per unit distance traversed by the particle. Therefore, a commonly used measure is the equivalent dose, which takes into account the type of radiation in terms of a radiation-weighting factor, which is then multiplied by the absorbed dose. The unit of the equivalent dose is sievert (Sv). Considering β-ray, X-ray, and γ-ray radiation, the equivalent dose is equal to the absorbed dose, whereas for α-ray radiation, the equivalent dose is assumed to be twenty times the absorbed dose (Cameron, 1991). For some applications gray and sievert are inconveniently large and thus milligray (mGy), defined as 1/1,000 gray, and millisievert (mSv), defined as 1/1,000 sievert, are frequently used instead.

Whole-body or significant partial-body exposure to ionizing radiation can increases the long-term risk for cancer (Brenner et al., 2003) and can cause the acute radiation syndrome. The onset and type of the symptoms of the acute radiation syndrome depends on the energy and dose of the exposure. Following radiation exposure, three distinct syndromes may occur: the cerebrovascular syndrome occurs after extremely high whole-body doses of radiation (> 30 Gy) as a result of hypotension and cerebral edema and is always fatal (Koenig et al., 2005). The gastrointestinal syndrome occurs after acute whole-body doses of approximately 6 to 20 Gy, primarily because of death of intestinal mucosal stem cells (Koenig et al., 2005). Typical symptoms of the prodromal phase (minutes to two days after exposure) are nausea and vomiting, headache, fatigue and fever (see Table 1.1 on the facing page). The third syndrome is the hematopoietic syndrome, which occurs after acute whole-body doses of approximately 2 to 10 Gy as a result of bone marrow depression. Lymphocyte depression can occur within 48 hour(s) (h). In Koenig et al. (2005) all three syndromes with their clinical signs, their medical treatment and the prognosis of exposed individuals are discussed in more detail.

1.3.2 Cytogenetic techniques for dose quantification

It has previously been shown that direct clinical signals, like the time to onset and the severity of typical symptoms of the acute radiation syndrome (see Table 1.1), can be correlated with the absorbed dose (Waselenko et al., 2004) and allow a rough estimation of the dose after whole-body acute exposures greater than 1 Gy (Simon et al., 2010). Based on these observations, several grading systems have been established,

	Dose range of acute whole-body exposure				
Symptoms and Signs	**Mild (1–2 Gy)**	**Moderate (2–4 Gy)**	**Severe (4–6 Gy)**	**Very severe (6–8 Gy)**	**Lethal (> 8 Gy)**
Vomiting					
- Onset	2 h	1–2 h	< 1 h	< 30 min	< 10 min
- Incidence	10–50%	70–90%	100%	100%	100%
Diarrhea			Mild	Heavy	Heavy
- Onset	2 h	1–2 h	3–8 h	1–3 h	< 1 h
- Incidence			< 10%	> 10%	100%
Headache	Slight	Mild	Moderate	Severe	Severe
- Onset			4–24 h	3–4 h	1–2 h
- Incidence			< 10%	> 10%	≈ 100%

Table 1.1: Symptoms of the prodromal phase of the acute radiation syndrome. The table is adapted from Koenig et al. (2005).

linking clinical symptoms with radiation dose and with prognostic probabilities of the patient's outcome (Dainiak et al., 2003; Fliedner et al., 2001). Similarly, haematological alterations, like the magnitude and rate of absolute lymphocyte depletion, can be used as a prognostic marker for estimating the radiation dose (Koenig et al., 2005). Both clinical signals and haematological alterations are useful for a fast, initial screening of exposed individuals, which is important to support prompt medical decision making and risk assessment, but their application is also limited for two reasons. First, they are not specific to ionizing radiation (Dainiak et al., 2003); pathologic agents, for example, can induce similar clinical signs and symptoms. Second, the minimum dose which can be estimated by such clinical and laboratory assays, is many times higher than the doses persons would be accidentally exposed to (Pinto et al., 2010). With lymphocyte depletion kinetics, for example, one can assess doses between 1 and 10 Gy with an exposure resolution of approximately 2 Gy (Waselenko et al., 2004). Thus, the sensitivity (*i.e.* the minimum detectable dose) of these techniques is too low for many practical applications.

Cytogenetic techniques allow for more accurate, precise and sensitive estimations of radiation doses. Ionizing radiation causes chromosomal aberrations which can serve as biomarkers for radiation exposure. For dose quantification, cytogenetic methods investigate either unstable chromosomal aberrations, such as dicentrics, centric rings and acentric fragments, whose persistence decline with the number of cell cycles, or stable chromosomal aberrations, such as translocations, which persist for a longer time period in the circulating lymphocytes. Three prominent cytogenetic techniques, each with their own strengths and weaknesses, are the dicentric chromosome assay, the cytokinesis-block micronucleus assay and the fluorescence *in situ* hybridization. These will be discussed in turn.

Dicentric chromosome assay: The dicentric chromosome assay has become the gold standard for dose quantification with a solid base of many years of research. The radiation dose is quantified by scoring dicentric chromosomes and centric rings in cultured lymphocytes and comparing their frequency to an *in vitro* dose-response (calibration) curve. The dose-response curve is produced by exposing blood specimens to doses of the appropriate quality and radiation rate.

The current gold standard is characterized by a high specificity and a good sensitivity to ionizing radiation. The minimal detectable dose following whole body exposure is approximately 0.1 to 0.01 Gy from the analysis of 500 to 1000 metaphase spreads for low and high LET radiation, respectively (Pinto et al., 2010). Radiation doses can be estimated also for partial body exposure through the application of mathematical procedures (Ainsbury et al., 2011). Nevertheless, this technique comes with several drawbacks: first, it is time-consuming and requires highly skilled technicians. Whereas the sample preparation takes at least 51 h, the subsequent data processing effort is 5 to 25 person hours per 500 cell analysis (Ainsbury et al., 2011). Second, dicentrics are unstable aberrations which are naturally eliminated from the PBL pool by apoptosis (Wojcik et al., 2004). This makes this technique inappropriate for the estimation of radiation doses of distant exposures. Especially, after high dose exposure, the dicentric frequency reduces rapidly due to lymphopenia with the consequence of a reduction in time from exposure to blood sampling. Third, irradiated cells may fail to arrive at metaphase for analysis due to a delayed cell cycle or cell death. Particulary in cases of partial body exposure, this may underestimate radiation doses (Pinto et al., 2010).

Cytokinesis-block micronucleus assay: Micronuclei are small extranuclear objects within the cytoplasm enclosing aberrated chromosome fragments or lacking chromosomes that did not properly segregate during cell mitosis. Similar to the dicentric chromosome assay, the number of micronuclei and their reference to an *in vitro* calibration curve allows for an estimation of radiation doses.

Micronuclei, which also belong to the unstable chromosomal aberrations, can arise from various clastogenic and aneugenic agents, and are thus not specific to ionizing radiation (Vral et al., 2011). It is also known that age and gender are confounding factors, inducing the frequency of micronuclei. The minimal detectable dose is approximately 0.2-0.3 Gy (Ainsbury et al., 2011). Compared to the dicentric chromosome assay, the cytokinesis-block micronucleus assay is easier to analyze, is less expansive and has a greater potential for an automated scoring process (Bolognesi et al., 2011). A major drawback is that the frequency of micronuclei is significantly lower than dicentrics, which implies that a substantially higher number of cells has to be analyzed in order to obtain comparable test accuracies. Further limitations are the inter-laboratory variability of dose–effect relationships for acute low LET radiation and the decrease of the anomalies with time (Pinto et al., 2010).

Fluorescence *in situ* hybridization: Fluorescence *in situ* hybridization is a staining technique which highlights different chromosomes in different colours by hybridizing fluorescent dye-labeled probes (Léonard et al., 2005). Additionally, the centromeres and telomeres of the chromosomes can be separately highlighted (multicolour fluorescence *in situ* hybridization). Hence, translocations, which are stable through mitosis, become visible as coloured rearrangements in a fluorescence microscope for scoring. Diagnostic systems based on stable chromosome aberrations are appropriate for retrospective biodosimetry, *i.e.* predicting long-term exposures. Drawbacks of analyzing translocations are that they are less specific to ionizing radiation than dicentrics and their frequency increases with age. Additionally, other confounding factors, like alcohol or nicotine, are known (Sigurdson et al., 2008).

1.3.3 Gene expression-based radiation dosimeters

Exposure to ionizing radiation leads to a complex genotoxic stress response, in which regulations at the transcriptional level play a central role (see Section 3.1 on page 28). In order to elucidate the underlying molecular mechanisms, researchers started to investigate the changes of gene expression in response to radiation exposure early. In 1992, Fornace and colleagues published a short list of mammalian DNA damage-inducible genes, most of them identified only a few years before (Fornace, 1992). The genes are associated to diverse cellular processes, such as signal transduction, response to tissue injury, DNA repair and response to oxidative stress, and thus already reflected the complexity of the transcriptional response. While researchers like Fornace could only investigate the expression levels of a limited number of genes, each of them separately measured, the advent of high-throughput technologies now allow the researcher to simultaneously monitor the expression of thousands of genes (see Section 1.2.1 on page 3). This technological milestone not only allows a deeper exploitation of the molecular responses following ionizing radiation, it also provides new opportunities for radiation biodosimetry: the information gathered about radiation-responsive genes can help to identify gene expression signatures, *i.e.* groups of potential biomarker genes, whose combined expression patterns reflect a specific transcriptional state and can be used for discriminating radiation doses. A bioinformatics-driven workflow to extract gene expression signatures from high-throughput data is presented in Section 1.2.2 on page 5. Note that in the context of radiation biodosimetry, the usage of supervised classification for biomarker discovery, as illustrated in Figure 1.2 on page 9 and as applied in this thesis, suggests to speak of radiation dose prediction, whereas biological-oriented biodosimetric studies tend to speak of dose estimation or dose quantification instead. I here employ both terms.

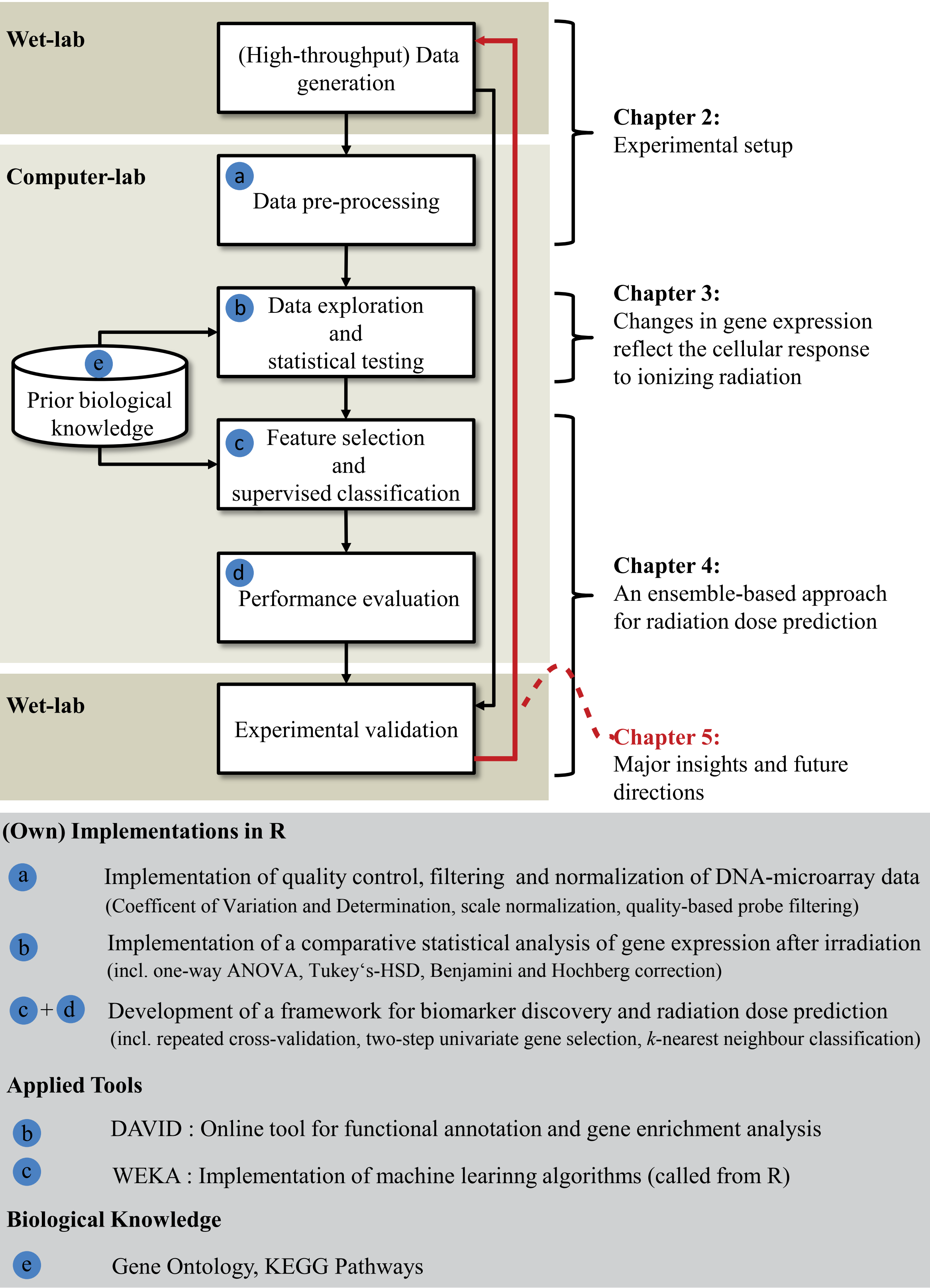
Amundson et al. (1999a,b) were among the first who investigated the correlation of radiation-induced gene expression patterns with radiation exposure by microarray-based gene expression profiling. With the aim of discovering new molecular biomarkers on the transcriptional level, they identified radiation-induced genes in human PBLs after *ex vivo* irradiation (Amundson et al., 2000). These initial results indicate that gene expression signatures may serve as radiation dosimeters allowing a fast, minimally invasive and automated screening of potentially exposed people. Brengues et al. (2010) recently developed a molecular bioassay with which *in vitro* irradiated blood samples can be distinguished from non-irradiated blood samples using an expression signature comprising of 14 reasonably chosen genes. This molecular bioassay, for which only a fingerstick of blood is needed, and data is delivered in less than 12 h, underpins the great potential and importance of molecular radiation biodosimeters.

Although the application of gene expression signatures as a tool for retrospective biodosimetry is promising, this recent branch will need many years of extensive research to become a standardized, validated method for biodosimetric applications. To this end, the knowledge of molecular biology and computational biology has to be integrated and, as proposed in this dissertation, the complimentary and synchronized research of both disciplines is important.

1.4 Outline of the thesis

This dissertation is structured into five chapters, including the introductory **Chapter 1** and the concluding **Chapter 5**. The latter summarizes the major findings of this thesis and gives a brief outlook to future research. **Chapter 2** offers a description of the experimental setup and techniques we used to monitor transcriptional changes in human PBLs after irradiation and provides therefore the basis for the results presented in the following chapters. In **Chapter 3**, I investigate the impact of radiation dose and time after exposure on the transcriptional response and derive biological implications by functionally characterizing the radiation-induced genes. In **Chapter 4**, a detailed description of the computational framework, I established for biomarker discovery and radiation dose prediction, is given. Its components, which I decided to incorporate, are discussed especially in the light of supporting biomarker reproducibility. Finally, the results obtained for the medium to high radiation doses as well as for the low radiation doses are presented and are set in the context of previously published gene expression-based biomarker studies. The appendix at the end of the thesis contains supplementary

information for chapter 4. A short synapsis at the beginning of each chapter outlines the content and sets the presented work in the overall context. It is additionally indicated on which publication(s) the work is based on. Already published work is reproduced with permission of the respective journals.
Figure 1.3 on the next page illustrates the outline of this work and additionally depicts how each chapter can be mapped to the components of the typical workflow for high-throughput data analysis and biomarker discovery presented above (see Figure 1.1 on page 4).


Wet-lab
(High-throughput) Data generation
Computer-lab
a Data pre-processing
b Data exploration and statistical testing
e Prior biological knowledge
c Feature selection and supervised classification
d Performance evaluation
Wet-lab
Experimental validation
Chapter 2:
Experimental setup
Chapter 3:
Changes in gene expression reflect the cellular response to ionizing radiation
Chapter 4:
An ensemble-based approach for radiation dose prediction
Chapter 5:
Major insights and future directions

(Own) Implementations in R

a Implementation of quality control, filtering and normalization of DNA-microarray data (Coefficent of Variation and Determination, scale normalization, quality-based probe filtering)

b Implementation of a comparative statistical analysis of gene expression after irradiation (incl. one-way ANOVA, Tukey‘s-HSD, Benjamini and Hochberg correction)

c + d Development of a framework for biomarker discovery and radiation dose prediction (incl. repeated cross-validation, two-step univariate gene selection, *k*-nearest neighbour classification)

Applied Tools

b DAVID : Online tool for functional annotation and gene enrichment analysis

c WEKA : Implementation of machine learinng algorithms (called from R)

Biological Knowledge

e Gene Ontology, KEGG Pathways

Figure 1.3: Illustrated outline of the thesis. All steps of the generic workflow for high-throughput data analysis and biomarker discovery explained in the Introduction and captured in Figure 1.1 can be mapped to chapters presented in this thesis. I realized most of the steps in R, a free language and environment for statistical computing and graphics (R Core Team, 2013), by combining own implemented functionalities with already existing functions of available add-on packages.

Chapter 2

Experimental setup

The experimental setup described in this chapter is based on the following publications:

- **Boldt S***, Knops K*, Kriehuber R, Wolkenhauer O (2012) A frequency-based gene selection method to identify robust biomarkers for radiation dose prediction. *International Journal of Radiation Biology* 3:267–276.
  *These authors contributed equally to this work.

- Knops K*, **Boldt S***, Wolkenhauer O, Kriehuber R (2012) Gene expression in low and high dose-irradiated human peripheral blood lymphocytes: Possible applications for biodosimetry. *Radiation Research* 178:304–312.
  *These authors contributed equally to this work.

Synopsis

Understanding the nature of experimental data is of crucial importance for establishing a reasonable analysis pipeline and for drawing biologically meaningful and trustworthy conclusions from the obtained results. In this chapter, I describe the experimental setup and experimental techniques used for monitoring radiation-induced gene expression changes in human PBLs. This is followed by a section explaining my strategy for quality assessment and pre-processing of DNA-microarray samples. The microarray gene expression measurements provide the basis for most of the results presented in this thesis.

2.1 Microarray-based gene expression profiling after irradiation

A valid statistical analysis of experimental data always requires a careful choice of experimental design. Ideally, biologists and data analysts jointly plan the experiments in order to achieve the objectives pursued by the study. For this reason, I was involved in devising the here described experimental setup from the outset, whereas the wet-lab experiments were carried out by Dr. Katja Knops at the Forschungszentrum Jülich.
To gain insight into the transcriptional response mediated by ionizing radiation, we monitored the gene expression of human PBLs after irradiation. We irradiated human PBLs with six radiation doses, ranging from low to high radiation doses, and measured their gene expression with DNA-microarrays at three different time points post exposure. The whole process, including gene expression profiling, quality assessment, data pre-processing and the validation of gene expression data by qRT-PCR is illustrated in Figure 2.1 on the next page. A brief description to each of these steps will now be given.

Irradiation of human peripheral blood lymphocytes: First, blood from a donor pool of three males and three females was obtained by venipuncture and the whole blood from each donor was collected in separate tubes. Afterwards, 9 millilitre (ml) aliquots of the heparinized blood were irradiated *ex vivo* using a Cs-137 γ-ray source at room temperature. Since we are interested in a comparative analysis of transcriptional changes after γ-exposure over a wide dose range, we decided to investigate six radiation doses, which can be categorized into low, medium and high radiation doses. For low radiation dose experiments the aliquots were irradiated with either 0.02 or 0.1 Gy at a dose rate of 0.0286 Gy/minute(s) (min), and for medium and high radiation dose experiments the aliquots were irradiated with 0.5 or 1 Gy and 2 or 4 Gy respectively at 0.7 Gy/min. In addition, time-matched non-irradiated control probes (*i.e.* aliquots irradiated with 0 Gy) were prepared for the low dose experiment as well as for the medium to high dose experiment.
Directly after irradiation, the lymphocytes were separated by density gradient centrifugation, which is a commonly used method for fractionating the whole blood into its different components through differential centrifugation and selective removal. After centrifugation, lymphocytes were found to be concentrated in a white layer between the plasma and the separation solution and were extracted.
To investigate not only dose-dependent but also time-dependent gene expression changes caused by ionizing radiation, the total RNA of the extracted lymphocytes were isolated at different time points after irradiation. For the low dose experiment, RNA was extracted at two time points, namely 24 and 48 h after irradiation, whereas for the medium and high dose experiment the RNA was extracted at three time points, namely 6, 24

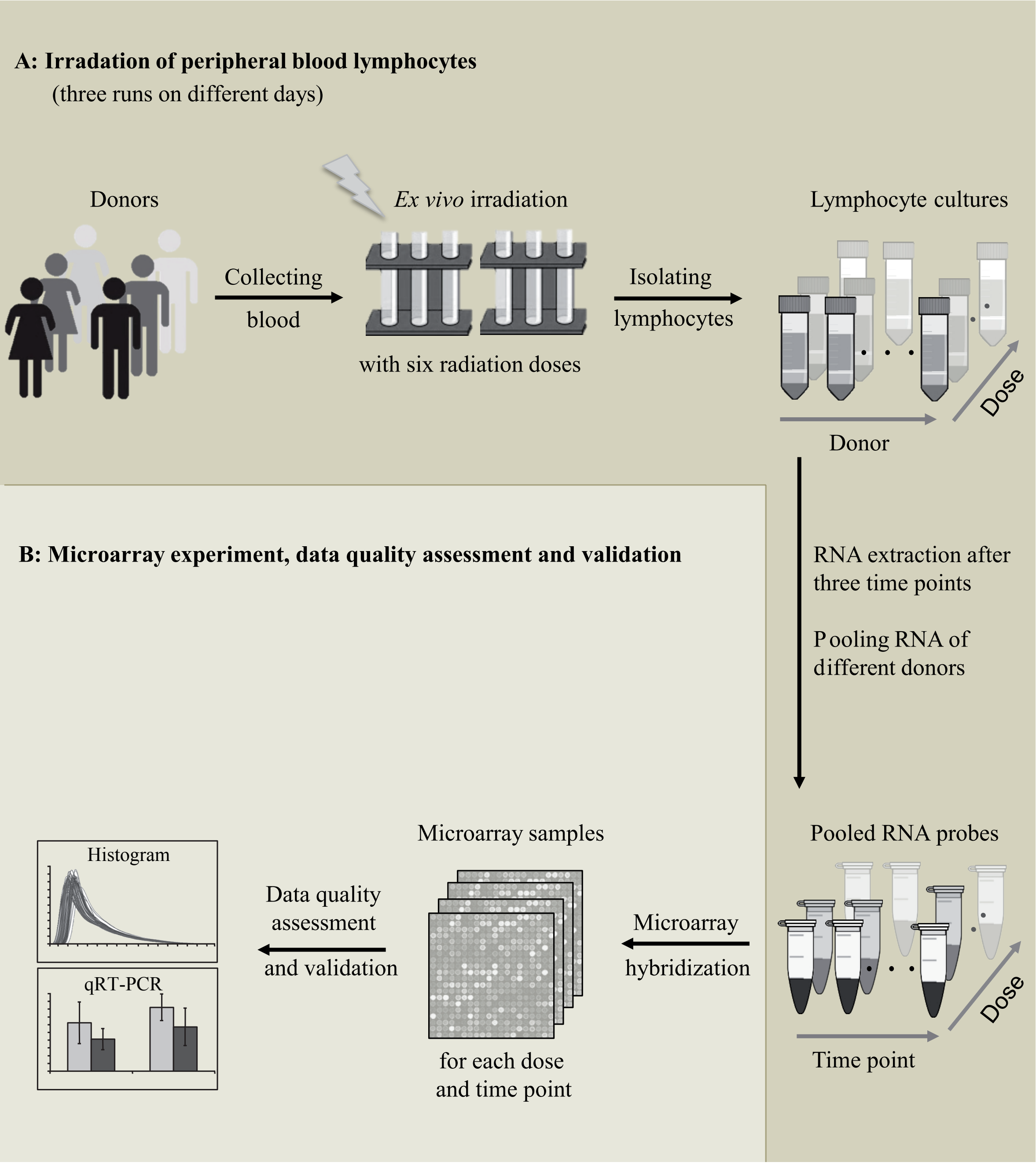


Figure 2.1: Experimental setup for irradiating human PBLs and the subsequent microarray experiment. Six healthy donors, three males and three females, donated blood, which was collected in separated tubes. Using a Cs-137 γ-ray source, aliquots of each tube were irradiated *ex vivo* with one of six different radiation doses: two low radiation doses (*i.e.* 0.02 and 0.1 Gy), two medium radiation doses (*i.e.* 0.5 and 1 Gy) and two high radiation doses (*i.e.* 2 and 4 Gy). Afterwards, lymphocytes were separated and for low radiation doses RNA was extracted 24 and 48 h after irradiation, whereas for medium and high radiation doses RNA was extracted 6, 24, and 48 h after irradiation. The total RNA from all donors isolated at the same time point and irradiated with the same radiation dose was pooled. This procedure was repeated three times, meaning that each donor donated blood three times on three different days, resulting in three independent experimental runs. The pooled RNA probes provided the basis for the subsequent microarray experiment, measuring the gene expression after irradiation. Finally, the quality of DNA-microarray samples was assessed and time-dependent and dose-dependent expression changes, measured by DNA-microarrays, were validated by qRT-PCR.

and 48 h after irradiation.

Finally, the total RNA from all donors isolated at the same time point and irradiated with the same radiation dose was pooled. The whole procedure, including blood collection, *ex vivo* irradiation, lymphocyte separation as well as RNA extraction and pooling was repeated three times, meaning that each donor donated blood three times on three different days. The pooled RNA probes of the three independent experimental runs provided the basis for the subsequent microarray experiment described in the next section.

DNA-microarray hybridization: Gene expression profiling can be used to measure the expression level of thousands of genes in parallel (see Section 1.2.1 on page 3). With DNA-microarrays, a common method for gene expression profiling, we studied the transcriptional response to ionizing radiation and, more specifically, the influence of the radiation dose and time elapsed since irradiation on the gene expression of human PBLs. According to the experimental procedure described in the previous section, we had the following experimental setting for the DNA-microarray experiment.

For the medium and high dose range, we had pooled RNA probes of human PBLs irradiated with five different radiation doses (0, 0.5, 1, 2 and 4 Gy) which were extracted at three different time points after irradiation (6, 24 and 48 h). Thus, we obtained 15 samples, each of them reflecting a different condition. By performing three independent experimental runs, we further obtained three biological replicates per condition, which finally resulted in 45 RNA samples for microarray hybridization. For the low dose range we had 18 pooled RNA probes, consisting of three biological replicates irradiated with three different radiation doses (0, 0.02 and 0.1 Gy) and extracted at two different time points after irradiation (24 and 48 h). For measuring the gene expression after irradiation we used one-color whole human genome microarrays from Agilent (Agilent Technologies, 4x44K, G4112F). These microarrays are manufactured by an *in situ* synthesis printing process, in which 60-mer oligonucleotides are deposited onto a glass side. They cover 41K unique human genes and transcripts.

Starting with our pooled RNA probes, each microarray experiment consisted of several consecutive steps: first, the mRNA of the pooled RNA probes was transcribed into cDNA, which was then further transcribed into complementary RNA (cRNA). Next, the cRNA was fluorescently labeled with Cyanine 3-CTP, and after cRNA purification, the dye incorporation and cRNA yields were measured. The labeled cRNA samples were coated onto the array and hybridized. Afterwards, the DNA-microarrays were washed and the slides were immediately scanned by which the fluorescence signal of each spot was detected. In a final step, the raw image data were further processed by image analysis and data extraction algorithms of the Agilent's Feature Extraction Software, which allowed a quantification of the detected signal intensities.

Validation of gene expression data by qRT-PCR: qRT-PCR is considered to be the gold standard technology for gene expression measurements because of its ability for accurate, sensitive and fast quantification of gene expression (Derveaux et al., 2010). It is common practice to validate the gene expression of a few candidate genes by qRT-PCR, since microarray data are known to be inherently noisy.

To validate and verify transcriptional changes induced after medium and high dose exposure, measured by DNA-microarrays, we performed qRT-PCR on (a) pooled RNA samples from the six healthy donors of the initial donor pool and (b) on non-pooled RNA samples of six healthy donors, of whom three donors belonged to the initial donor pool. The blood from which the RNA was isolated was irradiated *ex vivo* with 0, 0.5, 1, 2, and 4 Gy and isolated lymphocytes were cultured for 6, 24, and 48 h respectively. For the low dose range, qRT-PCR was performed only on the basis of non-pooled RNA samples. Therefore, blood of four healthy donors was irradiated *ex vivo* with 0, 0.02 and 0.1 Gy and lymphocytes were isolated immediately after irradiation. Following culturing, 24 and 48 h after irradiation, RNA was extracted from the lymphocytes.

The obtained fluorescence signals were normalized by the internal control dye (*i.e.* 5-Carboxy-X-Rhodamin) and the threshold cycle value (C_t) of the samples were normalized with respect to the C_t of the endogenous control glyceraldehyde-3-phosphate dehydrogenase 1 (GAPDH). Relative fold increase inductions were calculated by the delta delta cycle threshold ($\triangle\triangle C_t$) method. All samples were run in triplicates and gene expression was measured in three independent experiments.

2.2 Data quality assessment and pre-processing

Quality assessment and subsequent data pre-processing are important steps of each microarray analysis to ensure reliable and reproducible analysis results. Here, both steps were carried out separately on the samples measured at different times after irradiation.

I evaluated the quality of the DNA-microarray samples by comparison of intra- and inter-array replicates. To control the intra-array quality, the Coefficient of Variation (CV) of replicated non-control probes was calculated for each array (*i.e.* sample). The CV is defined as the ratio of the standard deviation to the mean and is often expressed as a percentage. When calculating the CV for each set of replicated probes, where signal variations can only arise from technical variations, a lower CV indicates a better reproducibility. Table 2.1 on the next page and Table 2.2 on page 25 report the obtained array CVs for all DNA-microarray samples irradiated with low radiation doses and medium to high radiation doses, respectively. The array CV is defined as the median of all CVs, calculated for 269 probes, each of them replicated 9 times. I compared the array CVs based on processed signal intensities (gProcessedSignal) and mean signal intensities (gMeanSignal) provided by the Agilent Feature Extraction Software Version. Since the array CVs obtained from the processed signal intensities were consistently lower than the CVs computed from the mean signal intensities, the processed signal intensities were used for all subsequent steps of our analysis. For inter-array quality assessment, I compared our biological replicates with the Coefficient of Determination (r^2), which denotes the strength of a correlation between two variables and thus measures to what portion the variation of one can be explained by, or attributed to, the other (Ding and Wilkins, 2004). Thus, identical DNA-microarray samples would have r^2 values of 1, and for an experiment without significant technical errors one would also expect high r^2 values between replicated arrays. As displayed in Table 2.3 on the next page the mean r^2 values for all pairs of biological replicates are between 0.93 and 0.99, indicating a high concordance for all replicated DNA-microarray samples.

		CV (%) - 24 h		**CV (%) - 48 h**	
Dose (Gy)	**Replicate**	**Processed Signal**	**Mean Signal**	**Processed Signal**	**Mean Signal**
0	1	0.67	1.75	0.61	1.31
0	2	0.56	1.68	0.56	1.58
0	3	0.61	1.68	0.52	1.6
0.02	1	0.55	1.64	0.58	1.51
0.02	2	0.6	1.65	0.61	1.68
0.02	3	0.62	1.71	0.58	1.68
0.1	1	0.57	1.75	0.56	1.67
0.1	2	0.6	1.75	0.58	1.55
0.1	3	0.72	1.7	0.62	1.71

Table 2.1: Assessment of the intra-array quality of DNA-microarray samples measuring transcriptional changes after low dose exposure. Using both the processed signal intensities (gProcessedSignal) and the mean signal intensities (gMeanSignal), the array Coefficients of Variation (CVs) were calculated for each sample of the low dose experiment (0, 0.02 and 0.1 Gy) by taking the median of the CVs obtained for all 269 replicated non-control probes. The array CVs obtained from the processed signal intensities were consistently lower than the array CVs computed from the mean signal intensities.

To sum up, in the comparison of intra- and inter-array replicates, no outliers (*i.e.* DNA-microarray samples, which are seriously affected by technical variations) could be detected and thus no samples had to be excluded from our dataset for further analysis. After quality assessment, I further pre-processed the processed signal intensities in the following way: initial data filtering was carried out to remove biased signal intensities which might hamper a valid statistical analysis and a biological interpretation of the data. Control features and nonuniform outliers, as well as signals that were flagged as not significantly above the background intensity in at least 25% of all samples, were therefore excluded. Afterwards, the processed signal intensities of the remaining probes were $\log_2$-transformed and subsequently median normalized, so that all DNA-

		CV (%) - 6 h		**CV (%) - 24 h**		**CV (%) - 48 h**	
Dose (Gy)	**Replicate**	**Processed Signal**	**Mean Signal**	**Processed Signal**	**Mean Signal**	**Processed Signal**	**Mean Signal**
0	1	0.73	1.71	0.72	1.78	0.73	1.67
0	2	0.96	2.17	0.66	1.54	0.77	1.6
0	3	0.64	1.72	0.64	1.65	0.71	1.61
0.5	1	0.68	1.44	0.93	1.44	0.68	1.65
0.5	2	0.78	1.92	0.62	1.61	0.99	1.43
0.5	3	0.67	1.57	0.89	1.6	0.72	1.59
1	1	0.75	1.8	0.66	1.66	0.67	1.6
1	2	0.57	1.32	0.74	1.66	0.7	1.48
1	3	0.86	1.7	0.7	1.66	0.93	1.71
2	1	0.69	1.42	0.71	1.79	0.71	1.51
2	2	0.71	1.73	0.71	2.08	0.7	1.8
2	3	0.69	1.73	0.67	1.66	0.93	1.99
4	1	1.08	1.86	0.93	1.73	0.92	1.64
4	2	0.76	1.74	0.71	1.62	0.96	1.84
4	3	0.71	1.93	0.71	1.77	0.95	1.98

Table 2.2: Assessment of the intra-array quality of DNA-microarray samples measuring transcriptional changes after medium and high dose exposure. Using both the processed signal intensities and the mean signal intensities, the array Coefficients of Variation (CVs) were calculated for each sample of the medium to high dose experiment (0, 0.5, 1, 2 and 4 Gy) by taking the median of the CVs obtained for all 269 replicated non-control probes. The array CVs obtained from the processed signal intensities were consistently lower than the array CVs computed from the mean signal intensities.

Experiment	**Dose (Gy)**	**Mean r^2**		
		6 h	**24 h**	**48 h**
Medium to high dose	0	0.99	0.99	0.95
	0.5	0.99	0.98	0.98
	1	0.99	0.98	0.95
	2	0.99	0.99	0.93
	4	0.99	0.99	0.97
Low dose	0	-	0.99	0.98
	0.02	-	0.99	0.93
	0.1	-	0.98	0.99

Table 2.3: Assessment of the inter-array quality of all replicated DNA-microarray samples. The mean Coefficient of Determination (r^2) for all replicates of the low dose experiment (0, 0.02, 0.1 Gy) and the medium to high dose experiment (0, 0.5, 1, 2, 4 Gy) range between 0.93 and 0.99, indicating a high concordance between replicated DNA-microarray samples.

microarray samples measured at the same time point after irradiation had the same median absolute deviation (Smyth and Speed, 2003).

Chapter 3

Changes in gene expression reflect the cellular response to ionizing radiation

The results presented in this chapter are published in the following publication:

Knops K*, **Boldt S***, Wolkenhauer O, Kriehuber R (2012) Gene expression in low and high dose-irradiated human peripheral blood lymphocytes: Possible applications for biodosimetry. *Radiation Research* 178:304-312
*These authors contributed equally to this work.

Synopsis

Radiation-induced DNA damage triggers a highly interwoven and cell-type specific network of intracellular and intercellular regulatory processes. Measuring the accompanying modulations at the transcriptional level allows researchers to characterize the involved functional processes and to elucidate the underlying molecular mechanisms. In particular, the biological effects of low radiation doses and their associated health risks in humans are as yet unclear. I therefore performed a systematic and comparative transcriptional analysis to investigate gene expression changes after low, medium, and high dose exposure in human PBLs. My analysis shows that both radiation dose and time after exposure have a substantial impact on the number of radiation-induced genes as well as on the affected pathways and molecular mechanisms. We further conclude that human PBLs show well-defined physiological responses even after acute low dose exposure.

3.1 The cellular effects of DNA damage

Ionizing radiation is a genotoxic insult which induces damage to the DNA either by direct or indirect interaction. In cases of direct interaction, the radiation particles directly collide with DNA, which can eventually cause DNA damage. If the radiation energy reacts with other molecules and the DNA damage is mediated by free radicals produced by that reaction one speaks of indirect interaction. Since water is a main cellular component, the radiolysis of water, *i.e.* the dissociation of water molecules by radiation, is the main reaction, contributing to the indirect interaction of ionizing radiation with DNA. Free radicals, if produced in vicinity of DNA, may diffuse and interact with DNA and thus cause DNA damage (Bajinskis, 2012). Whereas for high LET radiation, such as α and β-particles, the direct interaction is the dominant cause for harmful interaction with DNA, low LET radiation, including γ-rays, causes DNA damage mainly due to indirect interaction processes (Podgoršak, 2005).

Ionizing radiation induces a broad spectrum of DNA-lesions, including alterations of bases and sugars, protein-DNA and DNA-DNA cross-links, or strand breaks. The latter can be divided into single- and double-strand breaks where either only one or both complementary strands of the DNA double helix are broken. The inherent repair of a DNA double-strand break (DSB) is more difficult than that of other types of DNA damage and erroneous rejoining of broken DNA DSBs may lead to the loss, amplification or translocation of chromosomal parts (Khanna and Jackson, 2001). After exposure to ionizing radiation, the ionization events causing DNA damage are localized along the tracks of the ionizing particles generating clusters of ionizations. These clusters can induce multiple damages on both DNA strands in close proximity and thus generate DSBs with increased complexity. The complexity of DNA damage can have significant influence on the repair accuracy. However, the initiation of the multifarious DNA repair, which is broadly categorized into two complementary mechanisms, namely the error-free homologous recombination and the error-prone non-homologous end joining, is only one of several cellular processes triggered by DNA damage in order to protect the integrity of genetic information. The evolutionarily conserved, highly elaborate and complex network of intracellular and intercellular regulatory processes evoked by DNA damage is called the DNA damage response.

As described in Jackson (2002) and Khanna and Jackson (2001) the DNA damage response can be considered as a typical signal-transduction cascade in which DNA-lesions are physically detected by *sensor* proteins which then provoke the activation of *transducer* proteins to amplify and diversify the DNA damage signal by targeting a range of downstream *effectors* of the DNA damage response. Two proteins, namely ATM and ATR, play a dominant role in triggering different cellular responses following DNA damage in mammalian cells. Both kinases, stimulated in the aftermath of different types of DNA-lesions (Lagerwerf et al., 2011), activate a set of partly overlapping effector substrates, including p53 or CHK2, which lead to effects on cell-cycle progression, DNA repair and apoptosis (Jackson, 2002).

Since unrepaired DNA-lesions endanger the genomic integrity of dividing cells, the control of cell cycle progression in terms of initiating cell cycle checkpoints is of vital importance to prevent the replication of damaged DNA or segregation of damaged chromosomes. A prolonged cell cycle arrest provides the cell with time for the activation of DNA repair mechanisms (Khanna and Jackson, 2001) and thus provides an opportunity to monitor the appropriateness of cell death over DNA repair (Rich et al., 2000). Whenever DNA damage is mis- or unrepaired, cells can either enter a permanent cell cycle arrest which limits their proliferative competence (senescence) or the apoptotic programme which finally leads to cell death (Schmitt et al., 2007). Even when considered individually, DNA repair, cell cycle control and apoptosis, are very sophisticated processes of immense complexity, but when acting together in order to orchestrate the cellular response to DNA damage, they provide an immensely complicated, highly elaborated system, in which many aspects are still unclear. Underpinning its complexity, the DNA damage response swiftly modifies nearly every metabolic activity of the cell, including energy metabolism, cell−cell communication, and RNA processing (Rashi-Elkeles et al., 2011).

Radiation-induced DNA damage leads to significant gene expression modulations. The responses to DNA damage and therewith the changes at the transcriptional level hugely differ with the cell type. For example, splenic lymphocytes in a fetus and an adult readily initiate apoptosis after irradiation, but cardiac myocytes do not enter the apoptotic programme after exposure at any stage of the development (Rich et al., 2000). Measuring the gene expression by DNA-microarrays or next generation sequencing technologies accomplished by a statistical and bioinformatics-driven analysis provides the opportunity to investigate cell-specific responses induced by ionizing radiation. Especially, the molecular mechanisms triggered by low radiation doses are still unclear and require further investigations.

In this chapter, the results of a thorough statistical and functional analysis of transcriptional changes in PBLs after *ex vivo* $\gamma-$irradiation are presented. I statistically investigated not only gene expression changes across a wide range of radiation doses (0.02-4 Gy), but also the impact of time on the radiation response. Therefore, we monitored and examined transcriptional modulations after low (0.02 and 0.1 Gy), medium (0.5 and 1 Gy) and high dose (0.5 and 1 Gy) exposure at three different time points post irradiation (24 and 48 h after low dose exposure; 6, 24 and 48 h after medium and high dose exposure). With my comparative statistical analysis we identified radiation-induced genes as well as pathways and biological processes associated with dose-dependent and time-dependent gene expression changes after irradiation.

3.2 Methods for statistical and functional data analysis

3.2.1 Statistical testing

To identify genes with significantly altered gene expression after irradiation and thus, to investigate the effect of different radiation doses on gene expression at a specific time point after exposure, I conducted one-way Analysis of Variance (ANOVA). This is an appropriate statistical method when, as in my case,

1. the dependent variable (*i.e.* the gene expression) is of quantitative type (*i.e.* measured) and
2. one independent factor of qualitative type (*i.e. radiation dose*) with at least three levels is considered.

The levels of the factor *radiation dose* divides our microarray dataset into sample groups, in the following also referred to as dose groups. A sample group includes all biological replicates, irradiated with the same radiation dose and measured at a specific time point after exposure. Thus, I considered three sample groups for the low dose experiment (0, 0.02, 0.1 Gy) and five sample groups for the medium to high dose experiment (0, 0.5, 1, 2, 4 Gy).

One-way ANOVA determines whether there are any significant differences between mean expressions of all sample groups for a particular gene g. More precisely, the following null hypothesis is tested:

$$H_0 : \mu_i = \mu_l, \quad \forall \quad i, l = 1, ..., k \quad \text{and} \quad i \neq l, \tag{3.1}$$

where μ_i is the expectation of expression of gene g after irradiation with radiation dose i. k represents the number of radiation doses under study. μ_i is estimated by $\overline{x}_i$ which is defined as the mean expression of gene g based on all replicates irradiated with radiation dose i. Since the null hypothesis states that the expectations of all k sample groups are equal, the alternative hypothesis expresses that at least one of them differs from the others.

The rationale behind one-way ANOVA is that the total variability (SS_{total}) of the expression values of gene g can be partitioned into the between-group variability ($SS_{between}$), representing variability caused by factor *radiation dose*, and the within-group variability (SS_{within}), representing variability caused by chance (Ennos, 2007):

$$\begin{aligned} SS_{total} &= SS_{within} &&+ SS_{between}, \\ \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x})^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2 &&+ \sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x})^2, \end{aligned}$$

where x_{ij} is the expression value of gene g of replicate j ($j = 1, ..., n_i$) irradiated with radiation dose i ($i = 1, ..., k$) and $\overline{x}$ is the overall mean of all expression values measured for gene g. Thus, $SS_{between}$ is defined as the sum of squared distances of dose group mean from overall mean, weighted by sample size n_i, whereas SS_{within} is the sum of squared distances of individual expression values from dose group mean. By dividing $SS_{between}$ and SS_{within} by the appropriate degrees of freedom, one obtains the mean sum of squares for the variability of factor *radiation dose* ($MS_{between}$) and the mean sum of squares for the variability within dose groups (MS_{within}), where n denotes the total number of expression values of gene g:

$$MS_{between} = \frac{SS_{between}}{k-1},$$

$$MS_{within} = \frac{SS_{within}}{n-k}.$$

Based on this, the F-statistic is given by:

$$F = \frac{MS_{between}}{MS_{within}}. \tag{3.2}$$

Note that the larger the F value - illustrating that a greater part of variability of gene expression could be explained by the influence of factor *radiation dose* than by chance - the more likely is the rejection of the null hypothesis (see Equation 3.1 on the facing page) (Ennos, 2007).

Data to be analyzed with one-way ANOVA has to meet two main assumptions: first, the dependent variable has to be approximately normally distributed for each sample group. Second, the variance between all sample groups has to be the same. Whether these assumptions hold for our gene expression data was assessed by statistical tests for normality (*i.e.* Kolmogorov-Smirnov Test) and for homogeneity of variance (*i.e.* Levene Test).

After applying one-way ANOVA to all genes measured at a specific time point after exposure, I had to address the problem of multiple testing. Therefore, I adjusted the p-values with the method of Benjamini and Hochberg (1995) to control the false discovery rate (FDR) (*i.e.* the expected proportion of incorrectly rejected null hypotheses among all rejected hypotheses).

By means of one-way ANOVA, I identified genes with significant gene expression changes after irradiation at a specific time point after exposure. However, conducting one-way ANOVA does not answer the question of which particular radiation dose causes the significant change in gene expression. It is therefore necessary to additionally conduct a post-hoc-test for multiple comparisons of means. Thus, I finalized the statistical analysis by locating the pairwise differences in mean gene expression between dose groups by applying the Tukey's Honestly Significant Difference (Tukey's HSD) test (Tukey, 1949).

3.2.2 Functional enrichment analysis

To extract major biological meanings and to derive functional implications of the genes showing a significant altered gene expression after irradiation, we further performed a functional annotation and enrichment analysis. This analysis is based on the structured and controlled vocabulary provided by the GO Consortium for describing molecular function, biological process and cellular component characteristics of gene products. By using the Database for Annotation, Visualization and Integrated Discovery (DAVID) 6.7 (Huang et al., 2009a,b), we first identified over-represented GO Biological Process terms. These terms describe biological objectives to which a gene or gene product contributes (Ashburner et al., 2000). The GO-terms are structured in direct acyclic graphs, such that the hierarchical level of each term corresponds to the level of term specificity. In order to obtain more precise information for our functional annotation we decided to extract terms on the fifth level of the GO tree structure (Al-Shahrour et al., 2004). Based on the p-value of the modified Fisher's exact test (EASE score) we examined the significance of the GO-term enrichments within our data. The EASE score corresponds to the probability of obtaining an equal or greater frequency of the GO-term when randomly picking genes from the whole human background from Agilent. Furthermore, we additionally assigned the radiation-induced genes to affected KEGG pathways (Kanehisa et al., 2012).

In summary, with my statistical analysis I identified a set of radiation-induced genes for each radiation dose and each time point, which we then examined for over-representations of specific functions and pathways. A thorough comparative analysis of the obtained results allowed us to investigate time-dependent and dose-dependent biological effects mediated by ionizing radiation.

3.3 Comparative analysis of time- and dose-dependent gene expression changes

One aim of this thesis is to identify radiation-responsive genes and to analyze their time-dependent and dose-dependent expression changes after *ex vivo* γ-irradiation by DNA-microarray analysis. According to our experimental setting we investigated three different time points (6, 24, and 48 h) and seven radiation doses, including non-irradiated control samples (0 Gy). We grouped the radiation doses into three dose ranges, namely the low dose range (0.02 and 0.1 Gy), the medium dose range (0.5 and 1 Gy), and the high dose range (2 and 4 Gy). With my comparative statistical analysis I identified 1709 genes with significant radiation-induced expression changes. As depicted in Figure 3.1 on the next page the number of differentially expressed genes increases with increasing dose and time after exposure and significantly more genes are upregulated than downregulated.

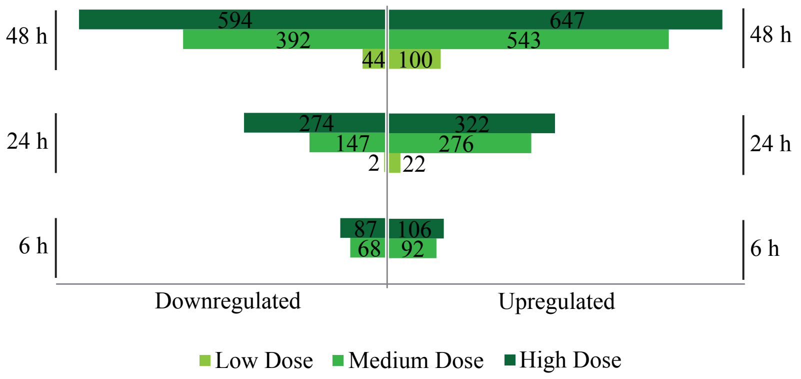


Figure 3.1: Results of the dose- and time-specific DNA-microarray analysis for the low, medium and high dose ranges. In the low (0.02 and 0.1 Gy), medium (0.5 and 1 Gy) and high dose range (2 and 4 Gy), the number of radiation-induced genes increased with increasing dose and time after exposure. At all examined radiation doses and time points after exposure, significantly more genes are upregulated than downregulated.

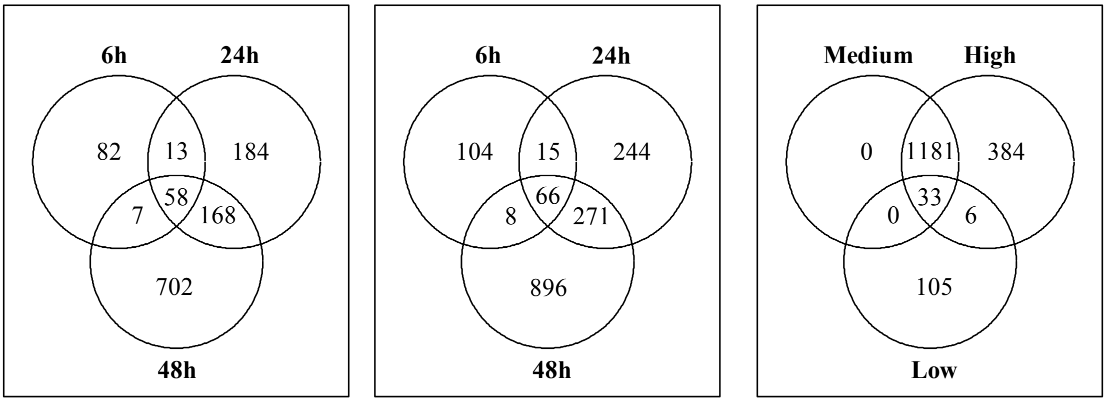
After low dose exposure I identified 144 differentially expressed genes 48 h after irradiation and no significant gene expression changes 24 h after irradiation. However, 24 genes exhibit at least a $\log_2$ fold-change of 1 and a variance smaller than the median variance across all low dose samples.

After medium dose exposure I observed 160 differentially expressed genes 6 h, 423 differentially expressed genes 24 h, and 935 differentially expressed genes 48 h after irradiation. The latter is more than a six-fold increase in the number of significant gene expression changes compared to the low dose at the same time point.

After high dose exposure I identified 193 differentially expressed genes 6 h and 596 differentially expressed genes 24 h after irradiation. With 1241 radiation-induced genes 48 h after high dose irradiation the number of significantly altered genes reaches a peak (see Figure 3.1).

Our findings are consistent with those reported by Jen and Cheung (2003) who assessed mRNA levels of genes in lymphoblastoid cells at various time points within 24 h following γ-irradiation and also reported a growing number of altered genes with increasing exposure dose. A conceivable explanation for this is that higher doses are known to produce more severe damage per cell (Pogosova-Agadjanyan et al., 2011), resulting in more radiation-induced expression alterations.

Next, I investigated the time-dependent overlap of genes with significant expression changes after medium dose as well as after high dose exposure. For the medium dose range, most of the genes show a significant expression alteration at only one of the examined time points: 82, 184 and 702 genes at 6, 24 and 48 h after irradiation. Only



(a) Medium dose range (b) High dose range (c) All dose ranges

Figure 3.2: Effects of radiation dose and time after exposure on the number of radiation-induced genes. The first two Venn diagrams show the time-dependent overlap of genes with significant expression changes 6, 24 and 48 h after irradiation **(a)** with medium radiation doses (0.5 and 1 Gy) and **(b)** with high radiation doses (2 and 4 Gy). The third Venn diagram **(c)** illustrates the dose-dependent overlap of all radiation-induced genes.

58 genes are differentially expressed at all time points (see Figure 3.2a). Similar results were obtained for the high dose range: half of the genes, showing an altered gene expression 24 h after irradiation (271 genes), are shared with the 48 h post-exposure times and only a total of 66 genes are differentially expressed at all examined times (see Figure 3.2b).

Finally, a comparison of radiation-induced genes after low, medium and high dose exposure revealed that all induced genes after medium dose exposure (1214 genes) are also induced after high dose exposure. Only 105 genes are exclusively differentially expressed after low dose exposure and 384 genes after high dose exposure (see Figure 3.2c).

3.4 Identification of radiation-induced processes and pathways

To examine biological processes and pathways affected by low to high radiation exposure, we functionally categorized the significantly altered genes with regard to the radiation dose and the time after exposure.

In the low dose range, the biological processes proteolysis and positive as well as negative regulation of apoptosis are significantly affected. This corresponds to the finding that the apoptosis rate is already increased 24 h after irradiation with 0.02 Gy (Knops, 2013). In a related study, published by Fachin et al. (2007), in which human lymphocytes were irradiated with two low radiation doses (0.1 and 0.25 Gy) and one medium radiation dose (0.5 Gy) and RNA was isolated for gene expression analysis at 48 h after stimulation, the main biological processes associated with modulated genes were metabolism, stress response/ DNA repair, cell growth/ differentiation and transcription regulation. A subsequent study, investigating individuals exposed to radiation (radia-

tion workers), whose low radiation doses ranged from 0.696 mSv to 39.088 mSv, showed an enrichment of several biological processes such as the ubiquitin cycle, DNA repair, cell cycle regulation/proliferation and stress response (Fachin et al., 2009).
After medium and high dose exposure, we observed two additional over-represented GO biological processes in response to radiation exposure, namely the nucleosome assembly and the DNA damage response (see Figure 3.3). Most altered genes are assigned to nucleosome assembly, the proteolysis and the regulation of apoptosis, which supports the known cell killing effect of ionizing radiation (Rich et al., 2000).
Altogether, samples exposed to low, medium and high radiation doses share 33 altered genes (see Figure 3.2c on the facing page). These are assigned to the regulation of the nucleobase, nucleoside, nucleotide and nucleic acid metabolic process and the cellular biosynthetic process. With a KEGG pathway enrichment analysis we further identified functional pathways which are associated with the time- and dose-dependent expression changes mediated by ionizing radiation (see Figure 3.4 on the next page). We detected three significantly enriched pathways after medium and high dose exposure: (i) the **p53 signaling pathway** (p-value 1.32×10^{-8} and 3.49×10^{-8} respectively), (ii) the **cytokine-cytokine receptor interaction pathway** (p-value 4.71×10^{-4} and 1.21×10^{-3} respectively), and (iii) the **systemic lupus erythematosus pathway**

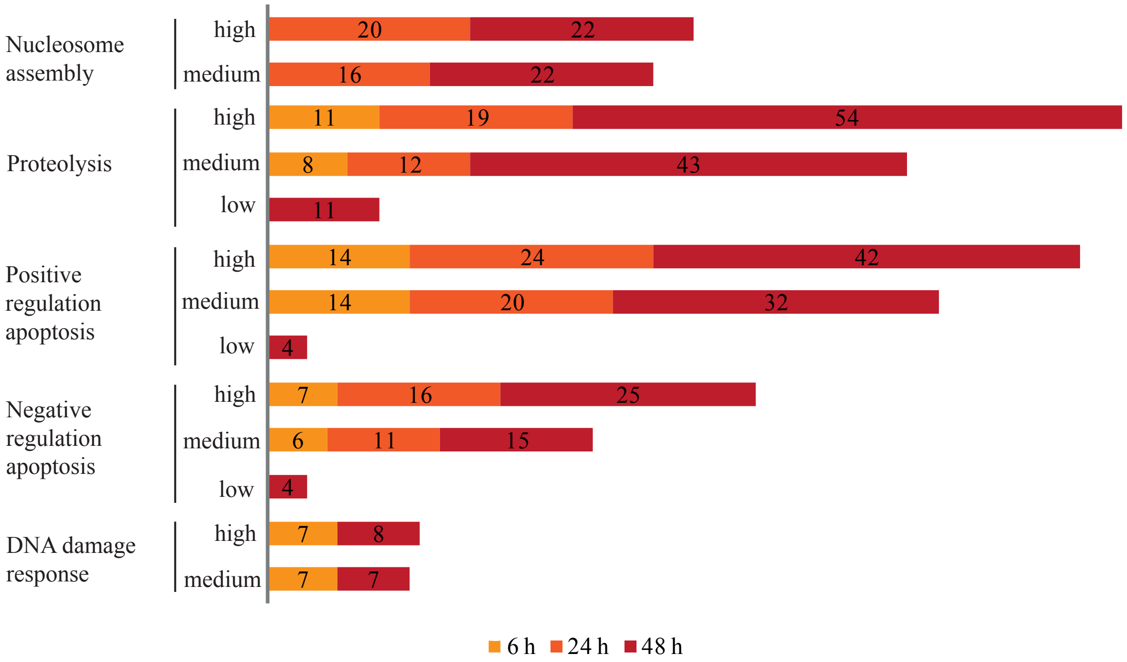


Figure 3.3: Enriched GO Biological Processes after low, medium and high dose exposure. The number of radiation-induced genes, assigned to different GO biological processes, increase with increasing dose and time after exposure. At all examined doses, significantly altered genes are involved in the proteolysis and in the positive and negative regulation of apoptosis.

(p-value 1.94×10^{-9} and 2.05×10^{-8} respectively). Remarkably, all genes that are assigned to these pathways after medium dose exposure are differentially expressed also after high dose exposure. In the following, I summarize and discuss our findings for each affected pathway separately.

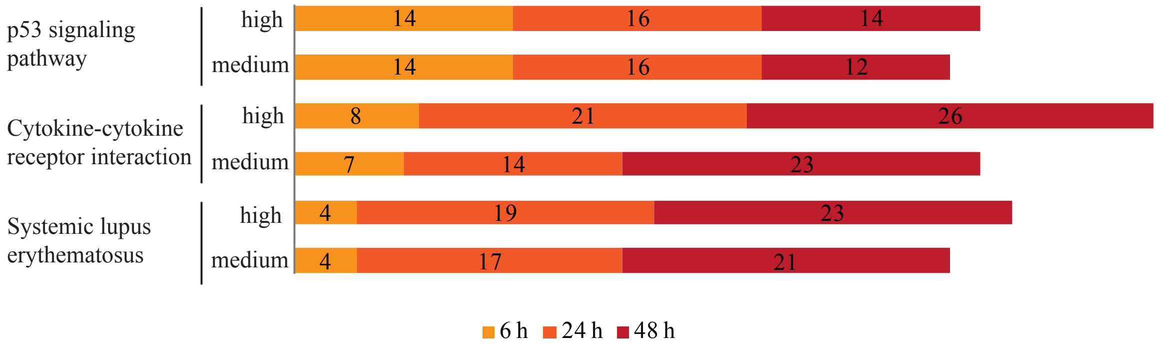


Figure 3.4: Enriched KEGG pathways after medium and high dose exposure. The number of radiation-induced genes involved in different pathways increase with increasing radiation dose and time after exposure. None of the significantly altered genes after low dose exposure (0.02 and 0.1 Gy) are assigned to a signaling pathway.

The p53 signaling pathway: p53 is a tumour suppressor protein that regulates the transcription of genes which are involved in a variety of cellular processes like apoptosis, cell cycle arrest and DNA repair. Being the major pathway activated by ionizing radiation, there is extensive literature concerning the role of p53 in the cellular response to radiation-induced DNA damage (Fei and El-Deiry, 2003; Helton and Chen, 2007; Lindsay et al., 2007; Nelson and Kastan, 1994).

Following radiation exposure, the transcription factor p53 is phosphorylated by ATM and DNA-PK and thereby activated. The ubiquitin E3 protein ligase MDM2, which promotes the degradation of p53 under normal cellular conditions, cannot bind to the phosphorylated p53 and hence the concentration of p53 increases (Campbell et al., 2013). After further post-translational modifications of p53 it transactivates numerous target genes and can thereby stimulate the DNA repair machinery and the activation of temporary and permanent cell cycle arrest, followed by down-stream cellular responses such as apoptosis to remove damaged cells (Budworth et al., 2012; Campbell et al., 2013). In line with the fact that more than 100 p53 downstream targets are currently known (Rashi-Elkeles et al., 2011), we also identified many differentially expressed genes regulated by p53 after medium and high dose exposure (see Table 3.1 on the facing page). These are involved in DNA repair, apoptosis and cell cycle control. Our functional analysis revealed that already 6 h after irradiation the p53 signaling pathway is strongly affected. At this time point 14 genes of the p53 signaling pathway

	6 h		**24 h**		**48 h**	
Gene Symbol	**Medium**	**High**	**Medium**	**High**	**Medium**	**High**
ATM						↓
BAX	↑	↑	↑	↑	↑	↑
BBC3	↑	↑	↑	↑	↑	↑
CDKN1A	↑	↑	↑	↑	↑	↑
CCND1	↑	↑				
CCNG1	↑	↑	↑	↑	↑	↑
CCNG2			↓	↓		
DDB2	↑	↑	↑	↑	↑	↑
EI24			↑	↑		
FAS	↑	↑	↑	↑	↑	↑
GADD45A	↑	↑	↑	↑	↑	↑
IGFBP3						↓
PIDD			↑	↑		
MDM2	↑	↑	↑	↑	↑	↑
PERP	↑	↑				
PPM1D	↑	↑	↑	↑		
SESN1	↑	↑	↑	↑	↑	↑
SESN2	↑	↑	↑	↑		
TNFRSF10B	↑	↑	↑	↑	↑	↑
TP53I3					↑	↑
ZMAT3			↑	↑	↑	↑

Table 3.1: Genes of the p53 signaling pathway with significant expression changes after medium or high dose exposure. Up arrows indicate that a gene is significantly upregulated after medium (0.5 and 1 Gy) or high dose (2 and 4 Gy) exposure, whereas a down arrow indicate a significant downregualation.

show an altered gene expression. Interestingly, the number of radiation-induced genes associated with the p53 signaling pathway remains nearly constant with increasing dose and time. With 16 p53 target genes, we detected the maximum number of genes linked to this pathway 24 h after exposure. 48 h after irradiation we identified 12 and 14 p53 target genes for the medium and high dose respectively (see also Figure 3.4 on page 36). Our results, showing that ionizing radiation has a strong effect on the transcriptional regulation of the p53 signaling pathway, are concordant with data from previous studies. For example, Mori et al. (2005) characterized the response of primary CD4$^+$ T lymphocytes to ionizing radiation by gene expression profiling and concluded that the majority of the strongly activated genes were p53 targets involved in DNA repair and apoptosis. In addition, Rashi-Elkeles et al. (2011) confirmed the central role of p53 in the transcriptional modulation induced by ionizing radiation by comparing the responses of cancerous and non-cancerous human cell lines with gene expression meta-analysis. Supporting the fact that gene expression varies widely in response to ionizing radiation in cell lines of different lineages (Meador et al., 2011), they observed a clear cell line-specific effect of radiation exposure. However, Rashi-Elkeles et al. (2011) additionally identified a set of genes which were common to the different cell lines, with the induced ones consisting almost exclusively of validated p53 targets, many of them also identified by our functional analysis (see Figure 3.1 on the preceding page).

The cytokine-cytokine receptor interaction pathway: Cytokines and their corresponding receptors, located on the cell surface, are involved in intercellular signal transduction and regulate biological processes like cell growth, differentiation, apoptosis or DNA repair. Our functional analysis revealed that the number of radiation-induced genes associated with the KEGG cytokine-cytokine receptor interaction pathway increases with increasing radiation dose and time after exposure. After medium dose exposure, seven genes of this pathway are differentially expressed 6 h after irradiation, 14 genes are differentially expressed 24 h after irradiation, and 23 genes are differentially expressed 48 h after irradiation. After high dose exposure, eight genes of this pathway are differentially expressed 6 h after irradiation, 21 genes are differentially expressed 24 h after irradiation, and 26 genes are differentially expressed 48 h after irradiation (see Figure 3.4 on page 36 and Table 3.2 on the facing page).

The systemic lupus erythematosus pathway: Systemic lupus erythematosus is a chronic autoimmune disease that culminates in the production of autoantibodies reactive with intracellular particles, consisting of nucleic acids and nucleic acid binding proteins (Kirou et al., 2005). A closer examination of the radiation-induced genes associated with the KEGG systemic lupus erythematosus pathway reveals that (a) the number of genes affected in this pathway increases with increasing time after exposure, whereas

	6 h		24 h		48 h	
Gene Symbol	**Medium**	**High**	**Medium**	**High**	**Medium**	**High**
CD27						↓
CD70	↑	↑	↑	↑	↑	↑
CD40	↑	↑				
CD40LG						↓
CCL5				↓	↓	↓
CCL17		↑				
CCL27				↑	↑	↑
CCR5			↑	↑		
CX3CR1				↓	↓	↓
CXCL2					↑	↑
CXCL3					↑	↑
CXCL16			↑	↑		
CXCR4				↓	↓	↓
EDAR						↓
FAS	↑	↑	↑	↑	↑	↑
FASLG				↓	↓	↓
IFNG					↓	↓
IFNGR2					↓	↓
IL21R			↑	↑		
IL18R1			↓	↓	↓	↓
IL18RAP				↓	↓	↓
IL2RB				↓	↓	↓
IL12RB2			↓	↓	↓	↓
OSM			↑	↑	↑	↑
PLEKHQ1						↑
TGFBR2						↓
TNFSF4	↑	↑	↑	↑	↑	↑
TNFSF8			↑	↑	↑	↑
TNFSF9	↑	↑				
TNFRSF17					↓	↓
TNFRSF10B	↑	↑	↑	↑	↑	↑
TNFRSF10C					↑	↑
TNFRSF10D			↑	↑		
TNFRSF13C	↓	↓				
XCL1			↓	↓	↓	↓
XCL2			↓	↓	↓	↓

Table 3.2: Genes of the cytokine-cytokine receptor interaction pathway with significantly altered genes expression after medium or high dose exposure. Up arrows indicate that a gene is significantly upregulated after medium (0.5 and 1 Gy) or high dose (2 and 4 Gy) exposure, whereas a down arrow indicate a significant downregualation.

the dose range seems to have only a marginal influence on the number of affected genes and (b) most of the affected genes are histone genes, showing an altered gene expression 24 h and 48 h after medium and high dose exposure (see Table 3.3 on the next page and Figure 3.4 on page 36). Histone proteins help to pack the DNA into ordered structures, called nucleosomes. In contrast to previous studies, our analysis revealed a radiation-induced upregulation of histone genes after medium and high dose irradiation in human PBLs. This finding differs from data published by Meador et al. (2011), reporting a negatively regulated histone gene expression in human lymphoblastoid and colon cancer cell lines. Additionally, the authors reported a radiation-induced cell cycle arrest in the S-phase of dividing cells, causing a halt of DNA synthesis and triggering histone gene downregulation, because no newly synthesized DNA had to be assembled by histones. Likewise, Su et al. (2004) described a histone gene downregulation after ionizing radiation by the dissociation of NPAT from histone gene promoters in a p53/p21-dependent manner, which resulted in inhibition of histone gene transcription. However, we monitored the gene expression of non-stimulated PBLs, which are in the G_0/G_1 phase and do not divide. Hence, the expression of histone genes is generally not required in these cells and is almost silenced. One explanation for the observed upregulation of histone genes after irradiation in PBLs might be that repair of radiation-induced DNA damage in G_0/G_1 requires newly synthesized histones for DNA packing, which leads to the observed upregulation of histone genes.

3.5 Summary of results

In the present chapter, I investigated the impact of radiation dose and time post exposure on gene expression modulations after ionizing radiation. To this end, I performed a systematic and comparative transcriptional analysis based on gene expression data of human PBLs measured 6, 24 and 48 h after medium (0.5 and 1 Gy) and high dose exposure (2 and 4 Gy), and 24 and 48 h after low dose exposure (0.02 and 0.1 Gy).
By conducting one-way ANOVA, a great number of genes exhibits significant radiation-induced gene expression changes. Underpinning the fact that higher radiation doses result in more severe damage per cell and consistent with the results reported by Jen and Cheung (2003), the number of genes increases with increasing dose and time after exposure, reaching a peak at 48 h after high dose exposure. An investigation of the time-dependent overlap of radiation-induced genes after medium dose as well as after high dose exposure revealed that most of the genes show a significant change in gene expression only at one of the examined time points. In contrast to that we observed a high dose-dependent overlap. All induced genes after medium dose exposure are also induced after high dose exposure. No genes show significant expression changes 24 h after low dose exposure which indicates that the gene expression modulations following

	6 h		24 h		48 h	
Gene Symbol	**Medium**	**High**	**Medium**	**High**	**Medium**	**High**
C3						↑
C7			↑	↑	↑	↑
CD40	↑	↑				
CD40LG						↓
FCGR2B	↓	↓				
HIST1H2AB			↑	↑	↑	↑
HIST1H2AD			↑	↑	↑	↑
HIST1H2AL			↑	↑	↑	↑
HIST1H2AJ			↑	↑	↑	↑
HIST1H2BB			↑	↑	↑	↑
HIST1H2BC			↑	↑	↑	↑
HIST1H2BD			↑	↑	↑	↑
HIST1H2BH			↑	↑	↑	↑
HIST1H2BJ			↑	↑	↑	↑
HIST1H2BL			↑	↑	↑	↑
HIST1H2BM				↑	↑	↑
HIST1H2BN				↑	↑	↑
HIST1H2BO			↑	↑	↑	↑
HIST1H4A			↑	↑	↑	↑
HIST2H2AB			↑	↑	↑	↑
HIST2H2BE			↑	↑	↑	↑
HIST3H2A			↑	↑	↑	↑
HIST3H2BB			↑	↑	↑	↑
HLA-DQA1	↑	↑				
HLA-DOB	↑	↑			↑	↑
IFNG					↓	↓

Table 3.3: Genes of the systemic lupus erythematosus pathway with significantly altered genes expression after medium or high dose exposure. Up arrows indicate that a gene is significantly upregulated after medium (0.5 and 1 Gy) or high dose (2 and 4 Gy) exposure, whereas a down arrow indicate a significant downregualation.

low dose exposure are not as pronounced.

We further identified biological processes and functional pathways, which are associated with the observed time-dependent and dose-dependent gene expression modulations after irradiation. The functional analysis revealed three pathways, namely the cytokine-cytokine receptor interaction pathway, the systemic lupus erythematosus pathways, and the p53 signaling pathway, which are significantly affected after medium and high dose exposure. Whereas the number of genes associated with the first two pathways increases with rising time, the number of genes associated with the p53 signaling pathway remains nearly constant with radiation dose and time post exposure. Our results obtained for the medium and high dose exposure are in line with similar studies investigating gene expression alterations after ionizing radiation (Jen and Cheung, 2003; Mori et al., 2005; Pogosova-Agadjanyan et al., 2011; Rashi-Elkeles et al., 2011). For the low dose range, we could not identify enriched biological pathways. However, we observed three enriched GO Biological Processes that are significantly over-represented already after low dose exposure: proteolysis, negative and positive regulation of apoptosis. Together with the finding that apoptosis is induced already 24 h after irradiation with 0.02 Gy (Knops, 2013), we conclude that acute low dose exposure, as low as 20 mGy, leads to well-defined physiological responses in human PBLs. With the here presented analysis we show that both radiation dose and time after exposure have a substantial impact on the number of radiation-induced genes, as well as on the affected pathways and molecular mechanisms in human PBLs.

Particularly in the light of our results demonstrating that the time after exposure influences the transcriptional response to ionizing radiation, the question arises of whether it is possible to identify a set of radiation-induced genes which allow an accurate retrospective estimation of radiation doses, regardless of the time post irradiation. This question is addressed in the next chapter in which I will introduce my computational and bioinformatics-driven framework for biomarker discovery and radiation dose prediction.

Chapter 4

An ensemble-based approach for radiation dose prediction

The results presented in this chapter are published in the following publications:

- **Boldt S***, Knops K*, Kriehuber R, Wolkenhauer O (2012) A frequency-based gene selection method to identify robust biomarkers for radiation dose prediction. *International Journal of Radiation Biology* 3:267–276.
  *These authors contributed equally to this work.

- Knops K*, **Boldt S***, Wolkenhauer O, Kriehuber R (2012) Gene expression in low and high dose-irradiated human peripheral blood lymphocytes: Possible applications for biodosimetry. *Radiation Research* 178:304–312.
  *These authors contributed equally to this work.

Synopsis

In the last chapter, it has been shown that radiation-induced gene expression changes are dependent on the radiation dose and the time after exposure. This finding raises the question of whether gene expression signatures allow for a, preferably time-independent, retrospective estimation of radiation doses. To address this question, I here present a computational and bioinformatics-driven framework which I developed and implemented for the discovery of potential biomarker signatures and the prediction of radiation doses. In the light of recent concerns about the reproducibility of molecular signatures identified for outcome prediction, the algorithmic design of my framework supports the identification of gene expression-based signatures, which are stable against small variations in the data. By applying my computational framework to our microarray data of irradiated human PBLs, I successfully extracted two candidate biomarker signatures with which low as well as medium to high radiation doses can be accurately assessed within a time frame that would be appropriate for medical decision making. To the best of my knowledge, the here presented work is the first gene expression study enabling a DNA-microarray-based dose prediction in the low dose range.

4.1 Research strategy

The lack of reproducibility of candidate biomarker signatures often impedes their transfer into clinical applications. It is therefore a fundamental challenge of current biomedical research to develop strategies to overcome this limitation. In this context, the concept of stable feature selection only recently gained importance in the field of computational biomarker discovery. For high-dimensional data, an important step towards the identification of potential biomarkers is to select promising features (*e.g.* genes) and to rank them according their relevance or importance. Based on such ranked lists, the final set of candidate biomarkers, like the top-k ranked features, is often determined. The term *stability* refers to the similarity of ranked lists obtained either by applying the same feature selection method to slightly modified versions of the underlying dataset (*e.g.* using different subsamples of the original dataset) or by applying different feature selection methods on the same dataset (Boulesteix and Slawski, 2009). High stability of features with respect to sampling variations is a good indicator for biomarker reproducibility (He and Yu, 2010), since markers which are tolerant against variations within data have a higher chance of having a high discriminatory power for experimental data from different studies generated in different laboratories.

It has previously been shown that genes which are selected for outcome prediction, are highly dependent on the training samples generated by a resampling strategy (Michiels et al., 2005). Different training samples often result in dissimilar signature genes which all allow an equally accurate outcome prediction (Ein-Dor et al., 2005). The existence of multiple sets of true markers is one of three main causes of instability (He and Yu, 2010). A second and important cause of instability is the small number of samples compared to the extremely high number of measured features in high-dimensional data (Ein-Dor et al., 2006; He and Yu, 2010). As demonstrated by Kim (2009), the overlap between independently developed gene signatures increases linearly with more samples. Based on a newly developed mathematical model, Ein-Dor et al. (2006) concluded that as a minimum, thousands of samples are needed to achieve a typical overlap of 50% between two predictive lists of genes obtained from breast cancer studies. As the third cause of low stability, He and Yu (2010) claim the application of algorithms that are primarily designed to select feature subsets providing the best prediction accuracy and do not explicitly attach importance to the stability.

Taking up the point of He and Yu (2010) that a reasonable algorithmic design for the selection of biomarker signatures is of crucial importance to support their stability, I developed an ensemble-based approach which incorporates a cross-validated univariate feature selection for the discovery of gene expression-based radiation dosimeters. As an integral part of my computational framework for biomarker discovery and radiation dose prediction, I assessed the classification performance of the extracted signatures in terms of their suitability for discriminating radiation doses. For the remainder of this section, the components of my computational framework, which I implemented in R, are described and discussed in more detail.

4.1.1 Components of the biomarker discovery framework

My computational framework for biomarker discovery and radiation dose prediction consists of four main components: 1. the **validation strategy** which implements a repeated cross-validation procedure, 2. the **gene selection** which incorporates a p-value-driven and fold-change-driven feature selection, 3. the **radiation dose prediction** for which a supervised classifier is trained, and finally 4. the **performance evaluation** of the classification process. In what follows, each component of my framework, designed for the identification of gene signatures to discriminate medium to high radiation doses, is explained separately. Figure 4.1 on the next page provides a schematic representation of my approach. To identify genes suitable for predicting low radiation doses, I established a modified version of the illustrated framework. The modifications are described in Section 4.1.3 on page 52.

Validation strategy:

A crucial aspect, which is often omitted in recent gene expression-based biodosimetric studies, is the need for a proper internal validation of a supervised classification. Methods, which pre-select potential biomarker genes out of samples that are again used to evaluate the prediction accuracy may lead to biased, overoptimistic classification results. It is therefore necessary that the samples used in the model building process are independent of the samples utilized for performance assessment (Baek et al., 2009). A common strategy is to construct training sets for model building and test sets for the subsequent performance assessment.

In order to split our data into training sets and test sets, I implemented a repeated stratified 9-fold cross-validation: the 45 pre-processed microarray samples, consisting of three biological replicates for each radiation dose (0, 0.5, 1, 2, 4 Gy) at each time point after irradiation (6, 24, 48 h), were randomly divided into nine parts in which each class, namely the radiation dose, is represented in the same proportion as in the full dataset. Each part, comprising five samples, was held out in turn as a test set in order to assess the accuracy of the trained supervised classifier. The remaining 40 samples built one training set. Based on these I extracted the potential biomarkers. I repeated the cross-validation procedure 100 times, meaning that the model building process of my computational framework, including the gene selection and classifier training, described below, operated on 900 training sets, whereas the 900 corresponding test sets remained unaffected until model validation, *i.e.* performance evaluation.

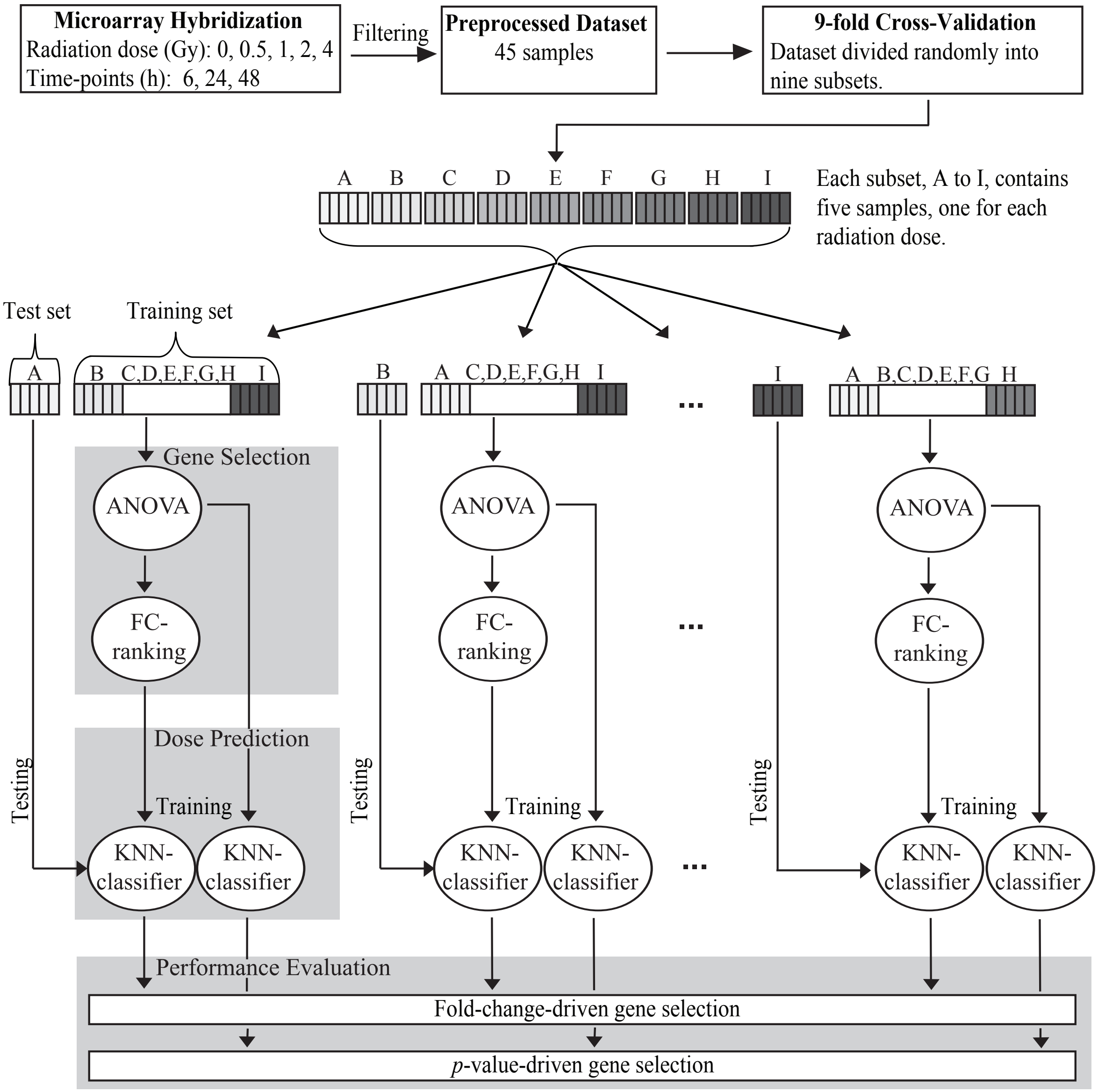


Figure 4.1: Schematic representation of my computational framework for biomarker discovery and radiation dose prediction. The main components of the framework for the identification of gene signatures for predicting medium to high radiation doses (0, 0.5, 1, 2, 4 Gy) are illustrated. The pre-processed dataset was randomly divided into nine parts, in which each class (*i.e.* radiation dose) is represented by the same proportion as in the full dataset. In turn, each part was held out once, functioning as a test set, whereas the remaining parts compose the corresponding training set. For each training set radiation-responsive genes were identified with one-way ANOVA. The expression patterns of each set of radiation-responsive genes were used to build k-nearest neighbour classifiers (KNN classifier), which were evaluated with the corresponding test sets. To obtain a final measure, the single performance results were averaged over all classifier predictions. Finally, ten genes with highest maximal fold-change between the measured radiation doses were selected from each set of radiation-responsive genes (FC-ranking). Again, these genes were used to build KNN classifiers, which were evaluated afterwards. The described procedure was repeated a hundred times and the obtained prediction performances of the two approaches were compared.

Gene selection:

In high-dimensional data the number of observations exceeds the number of samples by an order of magnitude. A selection of the most promising features is an indispensable step towards the identification of candidate biomarker signatures. Feature selection is of major importance because it usually enhances the prediction performance and reduces the risk of overfitting in the following classification process with supervised machine learning techniques. Assuming that the features selected for outcome prediction are also key-drivers of the biological system being studied, feature selection can also help to gain biological insights. Finally, the identification of a small signature of predictive features allow an inexpensive mass usage of custom-designed prognostic chips (Ein-Dor et al., 2006; Saeys et al., 2007). This issue is of particular importance in the work presented here. As I will discuss in Chapter 5 in more detail, a small set of expression-based radiation biodosimeters provide the basis for future biodosimetry devices applicable in cases of radiation accidents involving a large number of exposed individuals.
As already mentioned in the introduction (see Section 1.2 on page 3) one distinguishes between three categories of feature selection techniques with increasing degree of complexity (Saeys et al., 2007):

1. filter techniques select and rank features or feature subsets based on inherent characteristics of the underlying data. The most relevant or highest ranked features are subsequently used for classifier training.

2. wrapper techniques search for the best discriminating subsets of features by first selecting numerous subsets and then comparing them in terms of their prediction performance using supervised classification. Often the whole space of feature subsets is traversed using greedy search algorithms.

3. embedded techniques incorporate the feature selection process into the process of classifier training. Feature selection and classification are not separable.

Interestingly, there are very few studies investigating the impact of different feature selection methods on the performance and stability of biomarker signatures. Haury et al. (2011) empirically compared a panel of feature selection techniques from all three categories in terms of accuracy and stability. They demonstrated that filter methods, like a t-test-based feature selection, outperform more complex methods belonging to the category of wrapper and embedded techniques in terms of accuracy and stability. With the intention to assess the impact of different filter methods on prediction accuracy and signature stability, Fan et al. (2010) re-analyzed seven microarray cancer prognosis studies and performed a systematic parameter study. The authors showed that many more genes selected from repeated samplings were in common when a fold-change-based rule was used to select genes than when Pearson's correlation coefficient or a simple

t-test was applied. They concluded that the magnitude of differential expression with a loose p-value cut-off should be the preferred metric for gene selection, if stability and prediction accuracy is important. Drawing upon these findings, I decided to integrate a univariate filter technique into my computational framework for biomarker discovery and radiation dose prediction realized by a p-value-driven and fold-change-driven gene selection procedure.

p-value-driven gene selection method: For a gene expression-based discrimination of radiation doses it is a prerequisite that the utilized genes show an altered expression after irradiation. As already discussed in Section 3.2 on page 30 one-way ANOVA is an appropriate statistical method to identify genes with significant dose-dependent expression changes after radiation exposure. I here conducted one-way ANOVA for each training set constructed by the cross-validation procedure described above. By using *radiation dose* as the independent factor, the training set was partitioned into five sample groups, one for each radiation dose (*i.e.* 0, 0.5, 1, 2, 4 Gy). Herby, each dose group was composed of eight samples, including biological replicates for 6, 24, and 48 h after irradiation. Note that one of the actual nine samples of each dose group was used to construct the test set and was therefore not part of the training set.
The results of my comparative transcriptional analysis presented in Chapter 3 indicate that not only the radiation dose but also the time after exposure influences the gene expression after ionizing radiation (see Section 3.3 on page 32). Thus, it might seem questionable to include samples of different time points after exposure to the same sample group. In accordance with our experimental design (see Chapter 2), *time after exposure* could have been treated as a second independent factor for ANOVA. In light of having two factors potentially influencing gene expression, two-way ANOVA would have been the method of choice. But in terms of small sample sizes to be expected for two-way ANOVA (*i.e.* three samples per sample group in the whole dataset and only two samples per sample group in each training set) and good prospects to identify genes with a dose-dependent response even with an one-way approach (see Section 3.3 on page 32), it seemed reasonable to apply one-way ANOVA which is accompanied by an acceptable sample size for training sets. Coming back to Equation 3.2 on page 31, I expected to obtain large F-values for genes, whose gene expression are significantly affected by the factor *radiation dose*. Note that a large F-value results from a high proportion of $MS_{between}$, representing high variability in gene expression, and a small proportion of MS_{within}, indicating a marginal unexplained variability, and thus reflects the envisaged result of identifying genes with dose-dependent responses to radiation.

Fold-change-driven gene selection method: With the second gene selection method I further reduced the size of each list of radiation-responsive genes obtained by the *p*-value-driven gene selection. Assuming that genes with high expression changes after ionizing radiation are particularly useful for radiation dose prediction, I selected a small number of differentially expressed genes displaying high fold-changes.

In the following, I refer to the expression value of gene g of the replicated sample j, irradiated with radiation dose i, and measured at time point t, as $x_{ij,t}$. Due to the given experimental design, I consider $t = 3$ time points, $j = 3$ replicates and $i = 5$ radiation doses. The first step of my fold-change-driven gene selection was to calculate the mean expression $\overline{x}_i$ for each gene g based on all replicates irradiated with radiation dose i:

$$\overline{x}_i = \frac{1}{9} \sum_{j=1}^{3} \sum_{t=1}^{3} (x_{ij,t}) \quad \forall\, i = 1, ..., 5.$$

Please note that here $\overline{x}_i$ is the mean expression based on all replicated samples measured at all three time points t after irradiation with radiation dose i, whereas in Section 3.2 on page 30 $\overline{x}_i$ refers to the mean expression based on all replicated samples irradiated with radiation dose i and measured at one specific time point t.

In the second step, I calculated the maximal fold-change $MaxFC$ for each gene g. $MaxFC$ is defined as the difference of the maximal $\log_2$ mean expression $\overline{x}_i$ to the minimal $\log_2$ mean expression $\overline{x}_i$. Thus, I obtained for each gene g the maximal logarithmic fold-change after irradiation as follows:

$$\begin{aligned} MaxFC &= \log_2(\max_i(\overline{x}_i)) - \log_2(\min_i(\overline{x}_i)), \\ &= \log_2 \frac{\max_i(\overline{x}_i)}{\min_i(\overline{x}_i)}. \end{aligned} \tag{4.1}$$

In the last step of my fold-change-driven gene selection, I selected from each list of radiation-responsive genes obtained by the *p*-value-driven gene selection the ten genes with highest $MaxFC$.

To sum up, after repeating the procedure described above for all lists of radiation-responsive genes, the fold-change-driven gene selection resulted in 900 new lists, each containing ten radiation-responsive genes with highest maximal fold-change $MaxFC$.

Radiation dose prediction:

For radiation dose prediction, I performed KNN classification. This method belongs to the category of instance-based learning algorithms, which classify new observations by simply comparing them to a training set of already known instances (Aha et al., 1991). The only effort made in the training phase is to store all known instances in memory. No decision rules are inferred by generalizing the training dataset. Therefore, the KNN classifier is also called a *lazy* classifier. For classifier prediction, the class label of a new instance is predicted based on a majority vote among the class labels of its k closest training instances (*i.e.* nearest neighbours). The nearest neighbours are determined by a similarity measure, like the commonly used Euclidean distance function, which is defined as:

$$D(a,b) = \sum_{i=1}^{n}(a_i - b_i)^2, \tag{4.2}$$

where a and b are two instances and n is the number of numerical attributes each instance consists of. In addition to the choice of a similarity measure, the number of nearest neighbors k considered for majority voting can influence the prediction. When choosing a small number of nearest neighbours the classification is prone to inherent noise within the data, whereas a larger number of neighbours may lead to a less strict discrimination of class boundaries.

Since KNN classification is an intuitive method which can be also applied to small training datasets, I decided to integrate this obviously simple and intuitive instance-based learning method into my computational framework for biomarker and radiation dose prediction. More sophisticated methods like decision trees or support vector machines need a larger number of training instances to infer rules for classification and often additional parameters have to be determined or even optimized during the training phase.

How much the choice of classification method influences the prediction accuracy is a controversial issue. Whereas the study of Fan et al. (2010) demonstrated that the choice of classification methods minimally affect the prediction accuracy of microarray-based classifiers in cancer prognosis, Haury et al. (2011) observed that the best accuracy was achieved by the nearest centroids classifier, which is a method closely related to the KNN classifier. Besides the good performance results, these simple methods do not require an optimization of many classification parameters, making the computations fast and less prone to overfitting.

In the present scenario, each instance is represented by an expression profile, consisting of genes selected by one of my filter methods. In the training sets, the expression profiles are associated with their respective radiation doses, functioning as class labels, whereas in the test sets, the radiation dose is assumed to be unknown. For predicting the radiation dose, the k nearest neighbours were determined by the Euclidean distance measuring (see Equation 4.2), whereas the number of k, leading to an ideal classification result, was calculated by a nested cross-validation. Following this strategy, I performed radiation dose prediction by using (a) the 900 lists of radiation-responsive genes selected by the p-value-driven gene selection and (b) the 900 lists of ten genes with highest maximal fold-change identified by the p-value-driven and consecutively applied fold-change-driven gene selection.

Performance evaluation:

Only gene signatures which accurately predict the radiation doses under study are potential candidates for future radiation biodosimeters. Therefore, it is essential to systematically evaluate their performance in terms of dose prediction.
I calculated the average repeated 9-fold cross-validation performances for the p-value-driven and fold-change-driven gene selection method. The performance is expressed by three common performance measurements, namely the sensitivity, the specificity, and the overall success, which are defined as follows:

$$\begin{aligned} Sensitivity &= \frac{TP}{TP+FN}, \\ Specificity &= \frac{TN}{TN+FP}, \\ Overall\ success &= \frac{TP+TN}{TP+TN+FP+FN}. \end{aligned} \tag{4.3}$$

The true positives (TP) are the number of correctly predicted samples irradiated with class d (*i.e.* radiation dose i) and the true negatives (TN) are the number of correctly predicted non-d samples. The false negatives (FN) are defined as the number of incorrectly predicted samples of class d and the false positives (FP) are defined as the number of incorrectly predicted non-d samples.

4.1.2 Ensemble-based consensus signatures

In the beginning of this section, I introduced the term *stability* as a preferable characteristic of biomarker signatures, which is closely related to biomarker reproducibility. With my gene selection strategy, embedded in a cross-validation procedure, I established the basis for an approach referred to as ensemble-based feature selection. First, for each sample subset of the whole dataset (*i.e.* training set), potential biomarker signatures are selected. Second, the obtained candidate signatures, which are likely to differ from each other because they originate from slightly varying sample subsets, are aggregated to one ensemble-based consensus output. It has previously been demonstrated that ensemble feature selection can support the stability of biomarker signatures (Meinshausen and Bühlmann, 2010). Recently, Piao et al. (2012) and Abeel et al. (2010) successfully applied this approach to cancer classification.
A very intuitive way to aggregate sets of potential biomarker signatures is to choose only those genes for the ensemble-based consensus output that are most frequently part of the different signatures obtained by data perturbation (*e.g.* cross-validation). Since the most frequently chosen genes across all training sets are presumed to be highly robust against data variations and most relevant to sample prediction (Baek et al.,

2009; Chen et al., 2007), I followed this strategy to support the reproducibility of genes chosen for radiation dose prediction. The potential sets of biomarkers are significant radiation-responsive genes with high expression modulations after irradiation. With the two consecutively applied gene selection methods, I obtained for each training set a lists of ten genes with highest maximal fold-change. The ensemble-based consensus signature contains only genes that are part of all these 900 lists and is thus defined as the intersection of all lists with ten genes. The advantage of this approach is that effects of noise within data are eliminated because only those genes which are consistently found to respond with significant transcriptional changes to ionizing radiation are selected (Davis et al., 2006). Figure 4.2 on the facing page schematically illustrates the general procedure to construct ensemble-based consensus signatures.

4.1.3 Framework modifications for predicting low radiation doses

In the case of a large-scale nuclear accident with a widespread release of radioactivity, national health authorities may be requested to determine the individual radiation dose and the associated long-term health risks of a large number of exposed individuals. Since the *in vivo* radiation dose detection limit of cytogenetic methods is rather high, a biodosimetry method based on gene expression profiles to estimate even the smallest radiation doses would be beneficial.

Since my computational framework for biomarker discovery and radiation dose prediction was initially designed for the identification of gene signatures allowing a discrimination of medium to high radiation doses, I adapted the framework to the structure and characteristics of the microarray dataset measuring the gene expression alterations induced by low dose exposure (0, 0.02, and 0.1 Gy). In the following, the general procedure with its incorporated modifications for the low dose measurements is summarized. As described above, the initial step of my computational framework was to partition our microarray dataset into training and test sets, whereas the samples of the test sets were solely used for performance evaluation and not integrated into the model building process. In contrast to the former version of my framework, where partitioning was realized by a 9-fold cross-validation, I here implemented a 6-fold cross-validation because the present microarray data consists of fewer samples. Again, the stratified cross-validation procedure, ensuring that each test set contains one sample from each radiation dose, was repeated 100 times. The previously explained p-value-driven gene selection was used to identify genes displaying dose-dependent expression patterns after low dose exposure in each training set by applying one-way ANOVA. Here, the adjustment of the p-values for multiple comparison correction was modified. Instead of the Bonferroni adjustment, the less conservative method of Benjamini and Hochberg (1995) for controlling the FDR was used. This modification accounted for the more

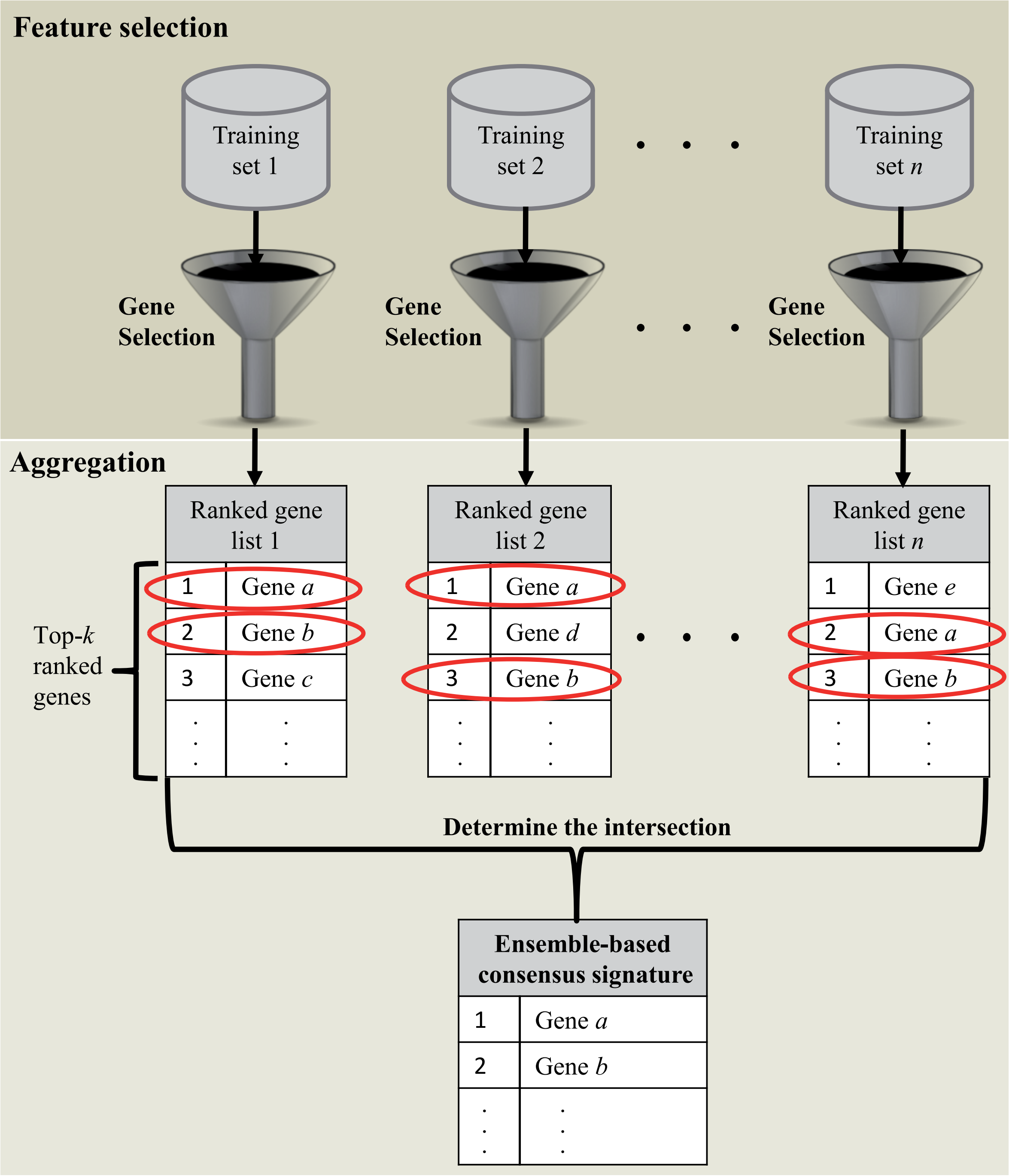


Figure 4.2: Schematic representation of the construction of an ensemble-based consensus signature. First, for each training set obtained by resampling, potential biomarker signatures are selected by applying a feature selection method. The selected signatures are likely to differ from each other because they originate from slightly varying sample subsets. Second, the signatures are aggregated to one ensemble-based consensus output. In the present work, I aggregated the expression-based signatures by determining their intersection.

subtle expression changes in the low dose range. Genes with an adjusted p-value < 0.05 were considered as significant radiation-responsive genes. With the consecutively applied fold-change-driven gene selection, I identified 20 genes displaying the highest fold-change after irradiation in each of the lists of radiation-responsive genes obtained by p-value-driven gene selection before. In this context, I slightly changed the computation of the fold-change (see Equation 4.1 on page 49) for the following reason: most genes show only subtle expression changes between 0.02 Gy and 0.1 Gy exposure which might impede a successful prediction. In order to support an accurate discrimination of both doses, I selected those genes which exhibit the greatest expression changes between 0.02 Gy and 0.1 Gy. Thus, the fold-change was defined as the difference of the mean log_2 expression values of the samples irradiated with 0.02 Gy and 0.1 Gy, and not as the maximal difference between the mean log_2 expression values of all radiation doses under study. Finally, all ranked genes lists, identified by the p-value-driven and consecutively applied fold-change-driven gene selection, were aggregated to construct an ensemble-based consensus signature. I selected all genes which are present in at least 500 of the 600 lists to build a consensus signature for low radiation doses. In the last step, I compared the consecutively applied gene selection methods with the consensus signature regarding their ability of radiation dose prediction. To this end, I calculated the average performances as already described in Section 4.1.1 (see Equation 4.3 on page 51).

4.2 Predicting medium to high radiation doses

In case of a radiation accident with a large number of affected persons without physical dosimeters, a quick and reliable method for dose determination is required. One goal of this thesis is the identification of stable biomarker signatures allowing an accurate prediction of radiation doses. For this purpose, I developed the computational framework for biomarker discovery and radiation dose prediction described above. In the following, I present the results obtained by applying my framework to gene expression data of human PBLs irradiated with medium to high radiation doses (0, 0.5, 1, 2, 4 Gy). In particular, the candidate biomarker signatures identified and their respective performances in terms of radiation dose prediction are discussed.

4.2.1 Outcome of the *p*-value-driven and fold-change-driven gene selection

Biomarker signatures applicable to radiation biodosimetry consist of radiation-induced genes, which show a dose-specific response post irradiation. The first step towards the detection of such genes is my p-value-driven gene selection. By applying one-way ANOVA with factor *radiation dose* to our DNA-microarray data, I identified genes with significant gene expression changes between at least two radiation doses under study. Based on the 900 lists of significant radiation-responsive genes, I assessed the performances of the cross-validated KNN classification for predicting medium to high radiation doses.

Averaged over all radiation doses, the *p*-value-driven gene selection procedure for radiation dose prediction yielded a sensitivity of 73%, an overall success of 89.2% and a specificity of 93.3%. It is noteworthy that 100% of all non-irradiated samples and 91.3% of the samples irradiated with 4 Gy are correctly classified. The majority of misclassifications occurs when predicting the radiation doses 0.5 Gy, 1 Gy and 2 Gy. The worst classification performance is obtained by predicting the samples irradiated with 1 Gy. Only 29.3% of them are assigned to the correct radiation dose. Coherent with these results is the specificity of the classification. The prediction specificities of the non-irradiated and 4 Gy irradiated samples are 100% and 99.9% respectively. The lowest specificity with 86.3% is measured by predicting the samples irradiated with 1 Gy.

As stated above, the *p*-value-driven gene selection exhibits flaws in predicting samples irradiated with 0.5, 1, and 2 Gy. Whereas 0.5 Gy exposure causes no acute medical deficits, a 2 Gy exposure is both immunosuppressive and myelosuppressive and could cause important clinical sequelae requiring medical intervention (Dressman et al., 2007). This is why 2 Gy may considered to be an important radiation dose for medical decision making. In this regard, a sensitivity of 65.3% for the identification of samples irradiated with 2 Gy, as obtained by the *p*-value-driven gene selection method, may be unsatisfactory for practical biodosimetry. This lack of performance is eliminated by the second, fold-change-driven gene selection method.

With the objective to further reduce the number of radiation-responsive genes used for radiation dose prediction, I subsequently applied the fold-change-driven gene selection. Hereby, ten genes with highest maximal fold-change (see Equation 4.1 on page 49) from each of the 900 lists of radiation-responsive genes obtained by the *p*-value-driven gene selection were extracted. Using the fold-change-driven gene selection, 95.7% of all test samples are correctly predicted. This implies that my consecutively applied feature selection enhances the sensitivity of the prediction by 22.6% in comparison to the *p*-value-driven gene selection. In particular, the number of correctly predicted samples irradiated with 1 Gy significantly increases to 90.3%, which is an explicit performance enhancement of 61%.

In summary, the consecutively applied fold-change-driven gene selection yielded an averaged overall success of 98.3% (+9.1% compared to the performance of the *p*-value-driven selection method), and a specificity of 98.9% (+5.6%). The differences of the measured prediction performances of the classification based on the *p*-value-driven and fold-change-driven selection of potential biomarker genes are shown in Table 4.1 on the following page. It additionally depicts that the selection of genes with the highest maximal fold-change leads to a significant improvement in the predictive power of radiation dose classification, particularly with regard to the samples irradiated with 0.5, 1, and 2 Gy.

Method	Dose (Gy)	Sensitivity	Specificity	Overall Success
p-value-driven				
	0	100	100	100
	0.5	79.2	88.7	86.8
	1	29.3	86.3	74.9
	2	65.3	91.4	86.2
	4	91.3	99.9	98.2
	Avg	73	93.3	89.2
Fold-change-driven				
	0	100	100	100
	0.5	100	98.1	98.5
	1	90.3	97.5	96.1
	2	90.1	99.8	97.8
	4	97.9	99.2	98.9
	Avg	95.7	98.9	98.3
Consensus signature				
	0	100	100	100
	0.5	100	97.2	97.8
	1	68	99.6	93.3
	2	89	91.9	91.3
	4	88.4	97.6	95.8
	Avg	89.1	97.3	95.6

Table 4.1: Prediction performances obtained for the medium to high radiation doses. Compared are the performances yielded by the *p*-value-driven gene selection, the consecutively applied fold-change-driven gene selection and the ensemble-based consensus signature when predicting the radiation dose of samples after medium to high dose exposure (0, 0.5, 1, 2, 4 Gy). The sensitivities, specificities and overall successes obtained for each radiation dose and averaged over the complete dose range (Avg) are given in percent. This table is adapted from Boldt et al. (2012).

4.2.2 The consensus signature and its prediction performance

Across all 900 lists obtained by the *p*-value-driven and consecutively applied fold-change-driven gene selection 16 different genes were identified (*i.e.* union of all 900 lists). Consequently, all lists of radiation-responsive genes with the highest maximal fold-change consist of a varying composition of these 16 genes (see Table 4.2 on the next page).

Using the ensemble-based consensus signature, 89.1% of the test samples are correctly assigned to their radiation dose. Similar to the classification results obtained with the *p*-value-driven gene selection, the performance evaluation revealed that the prediction of samples irradiated with 1 Gy leads to the highest misclassification rate; 32% of all samples irradiated with 1 Gy are incorrectly predicted. In contrast, 100% of all test samples irradiated with 0 Gy and 0.5 Gy are correctly assigned to their radiation dose. Averaged over all radiation doses, the dose prediction with the ensemble-based consensus signature yielded an average specificity of 97.3% and average overall success of 95.6%. The detailed performance results obtained by all gene selection strategies are compared in Table 4.1 on the facing page. It additionally highlights that the consensus signature yielded a significantly increased predictive power in comparison to the p-value-driven gene selection, as well as a slightly decreased performance in comparison to the fold-change-driven gene selection. The genes of the ensemble-based consensus signature are assumed to have a high stability with respect to sample variations. Seven of the 16 genes, are part of all lists which I obtained by the p-value-driven and fold-change-driven gene selection and are thus part of the consensus signature. Namely these genes are: *TNFSF4*, *FDXR*, *DOK7*, *SPATA18*, *PHLDA3*, *LGR6*, and *VWCE*. The detailed function of the herein most frequently detected signature genes as well as their implication in radiation response is still not fully resolved. *FDXR* is a target gene of the p53 family that can be induced by DNA damage in cells in a p53-dependent manner (Liu and Chen, 2002). Kawase et al. (2009) recently reported that *PHLDA3* is a p53 target gene

Agilent ID	**Gene Symbol**	**Frequency of Selection**
A_23_P126836	*TNFSF4*	900
A_23_P38154	*FDXR*	900
A_23_P398854	*DOK7*	900
A_23_P407112	*SPATA18*	900
A_23_P62959	*PHLDA3*	900
A_24_P773539	*LGR6*	900
A_23_P52986	*VWCE*	900
A_23_P408285	*PRICKLE1*	674
A_32_P347617	*RP4-742C19.3*	599
A_32_P85230	*ISG20L1*	565
A_32_P170454	*LOC283454*	371
A_23_P357717	*TCL1A*	238
A_32_P156786	*THC2651023*	168
A_23_P84399	*CNTNAP2*	41
A_23_P31681	*C8orf38*	26
A_32_P210202	*E2F7*	18

Table 4.2: List of all radiation-responsive genes identified by the fold-change-driven gene selection for the medium to high dose range. For the medium to high dose range (0, 0.5, 1, 2, 4 Gy), 16 genes were selected by the p-value-driven and consecutively applied fold-change-driven gene selection. The frequency of selection indicates how often a gene is among the ten selected genes with maximal fold-change within 100 repeated 9-fold-cross-validations. Seven of the 16 genes are part of all 900 lists.

that has been implicated in apoptosis, but its definite role in radiation response remains yet to be determined (Amundson et al., 2008). As recently published by Bornstein et al. (2011), *SPATA18* is also a transcriptional target of p53. Ionizing radiation induces a large variety of DNA-lesions, including single-strand breaks and double-strand breaks, as well as base and sugar damage (Fei and El-Deiry, 2003). If DNA damage repair fails, it is very likely that *FDXR*, *PHLDA3* and *SPATA18* as p53 target genes are upregulated after irradiation, probably in the context of radiation-induced apoptosis. *TNFSF4* is the ligand for tumor necrosis factor receptor superfamily, member 4 and plays a role as an essential late co-stimulatory signal that is required to maintain long-term cluster of differentiation CD4$^+$ T-cell survival by suppression of apoptosis (Wang et al., 2009a). Therefore, *TNFSF4* might be upregulated after irradiation in order to prevent or counteract radiation-induced apoptosis in irradiated lymphocytes. Up until now, the function of *VWCE* and *DOK7* after irradiation is not be understood. *VWCE* is considered to be a regulatory element of the β-catenin signaling pathway and has been recently shown to activate β-catenin in tumor cells (Du et al., 2010). *DOK7* is essential for neuromuscular synaptogenesis and induces the tyrosine phosphorylation of muscle, skeletal, receptor tyrosine kinase (MuSK), resulting in numerous differentiated acetylcholine receptor (AChR) clusters (Vogt et al., 2009).

4.2.3 Successful experimental validation of microarray data

As previously described, I performed DNA-microarray gene expression analysis of irradiated PBLs to construct an ensemble-based consensus signature. This signature consists of seven genes, which were selected by the criteria of being among the genes with maximal fold-change in each of the 900 obtained lists of radiation-responsive genes (see Table 4.2 on the previous page). Based on the assumption that the seven genes are stable against slight variations in our gene expression data, we selected them for validating the radiation-induced gene expression alterations by qRT-PCR analysis. Therefore, pooled RNA samples of six healthy donors we previously used for DNA-microarrays and non-pooled RNA samples of six healthy donors, three of them belonging to the initial donor pool in combination with three additional donors, were applied for qRT-PCR analysis. Thereby, we detected very similar gene expression profiles for *TNFSF4*, *SPATA18* and *VWCE* at all examined time points and radiation doses in DNA-microarray and qRT-PCR measurements. *FDXR* and *PHLDA3* feature approximately the same gene expression profiles 6 h after irradiation, whereas 24 h and 48 h post exposure the gene expression changes measured by DNA-microarrays are slightly higher than the alterations detected by qRT-PCR in the pooled as well as in the non-pooled samples (see Figure 4.3 on the facing page and Figure 4.4 on page 60). For *DOK7* the qRT-PCR measurements show a stronger upregulation after irradiation in the non-pooled samples when compared with the pooled samples. *LGR6* exhibits only a low upregulation in the qRT-PCR measurements in comparison to the DNA-microarray data (data not shown).

As already described by Pogosova-Agadjanyan et al. (2011) and Kang et al. (2003) microarray and RT-PCR or qRT-PCR measurements can yield different expression results. In this study, qRT-PCR measurements of the low expressed gene *LGR6* do not show strong expression alterations after irradiation when compared to the microarray measurements. One reason for this could be that alternative splicing forms, as observed for 40-60% of human genes, hybridized to that particular probe on the microarrays (Rockett and Hellmann, 2004). Furthermore, genes with very high or low levels of expression showed reduced agreement between RT-PCR and microarray results, whereby the expressions were undetectable by RT-PCR but appreciable by microarrays (Etienne et al., 2004).

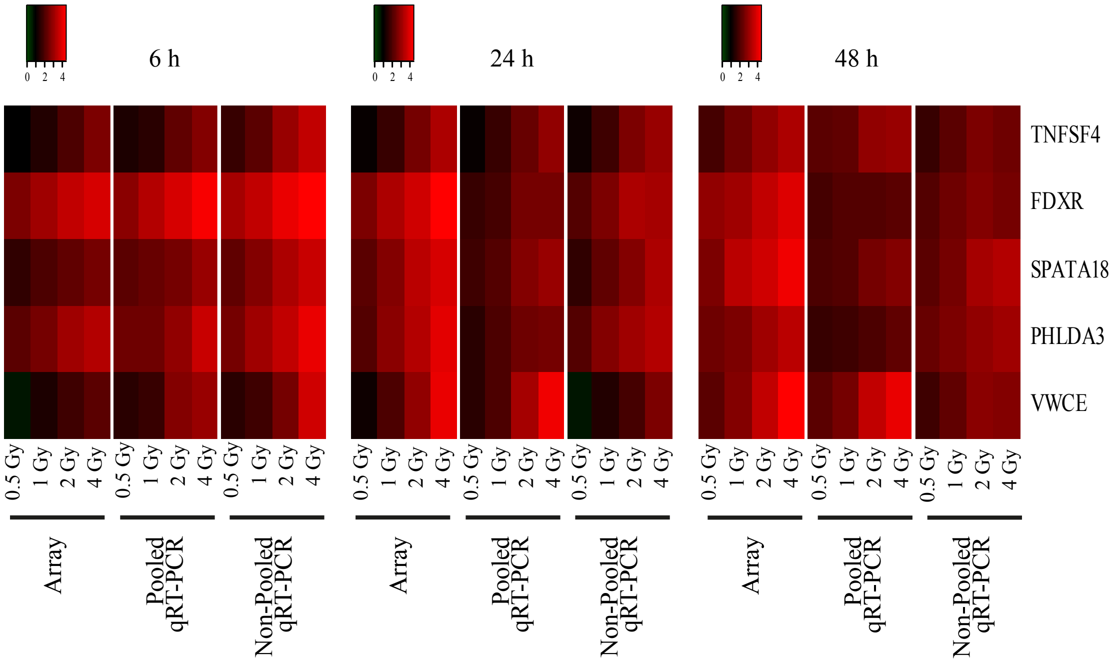


Figure 4.3: Heatmaps comparing radiation-induced expression changes of potential biomarkers obtained by DNA-microarray and qRT-PCR analysis. The $\log_2$ fold-changes of *TNFSF4*, *FDXR*, *SPATA18*, *PHLDA3*, and *VWCE* (rows) are selected to compare the expression changes obtained by DNA-microarray and qRT-PCR analysis of pooled and non-pooled samples (columns). Illustrated are the expression alterations after medium to high dose exposure (0.5, 1, 2, and 4 Gy) measured 6, 24, and 48 h after irradiation. High $\log_2$ fold-changes are depicted as red.

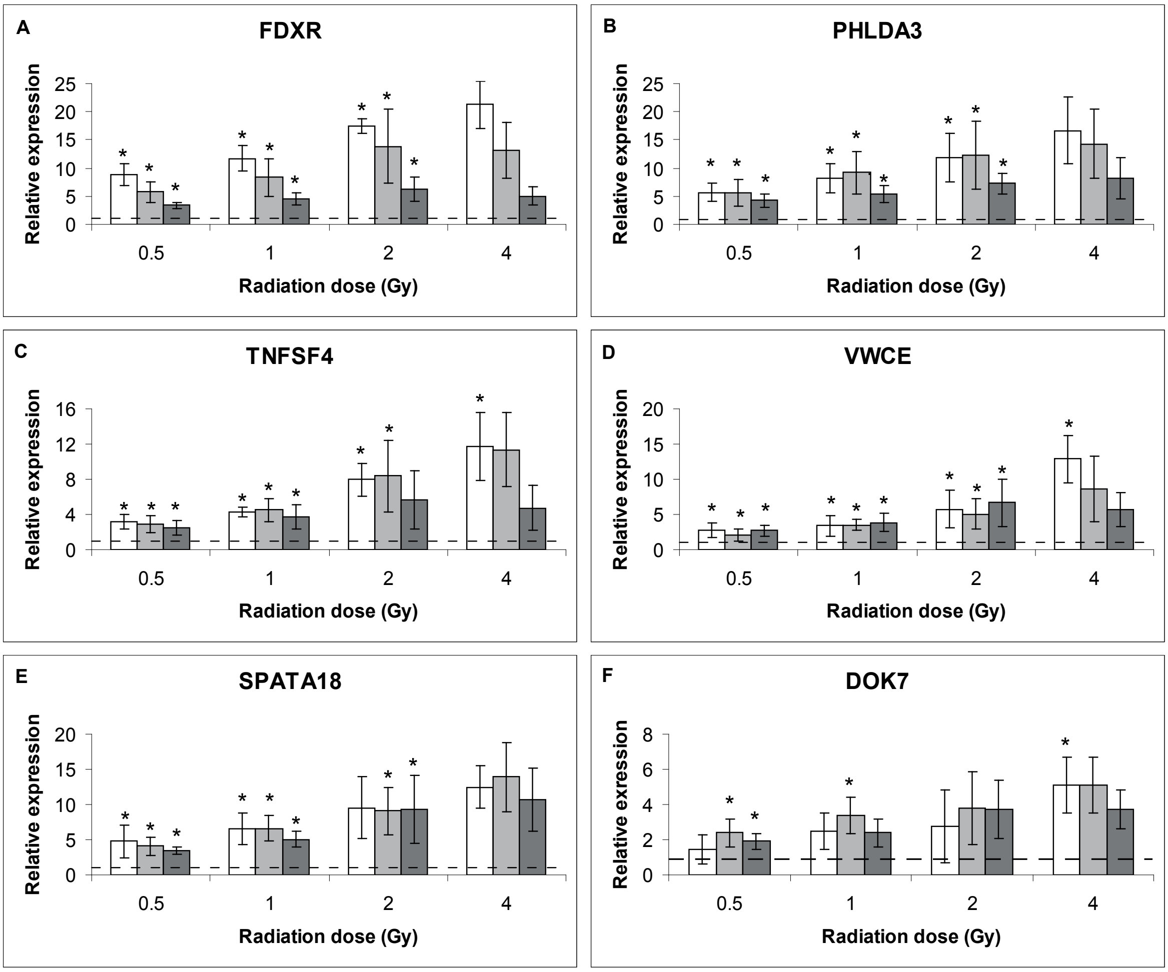


Figure 4.4: Illustration of the mean value of relative expression of six radiation-responsive genes in non-pooled, individual RNA samples based on qRT-PCR after medium and high dose exposure. The relative expressions 6 h (white bars; $n = 6$), 24 h (light grey bars; $n = 6$) and 48 h (dark grey bars; $n = 6$) post-irradiation measured by qRT-PCR based on the expression of the non-irradiated control (dashed line) are shown. A significant rise in the gene expression from one dose to the next higher dose is marked by an asterisk (p-value < 0.05). All of the displayed genes reveal a dose-dependent increase of the gene expression. All given data points are significantly different to the non-irradiated control ($p < 0.05$). **A:** *FDXR* exhibits a very strong (9- to 20-fold) upregulation 6 h post exposure. **B and C**: *PHLDA3* and *TNFSF4* show a similar gene expression profile at all examined time points post irradiation with a less pronounced increase at later time points. **D and E:** The expression of *VWCE* and *SPATA18* increases with increasing dose and seems to be virtually irrespective to the time post irradiation (with the exception at 4 Gy). **F:** *DOK7* shows the strongest upregulation 24 h after irradiation and is rather indicative in the lower dose range. Error bars indicate the standard error of the mean (SEM) for $n = 6$ independent experiments.

4.3 Predicting low radiation doses

A gene expression-based estimation of low radiation doses would be beneficial to assess long-term health risks associated with low dose exposure. In the present section, I discuss (1) whether samples irradiated with low radiation doses can be discriminated based on gene expression levels, and (2) whether my ensemble-based approach is a valid tool for identifying promising biomarker candidates for the low dose range (0, 0.02, 0.1 Gy). First, I identified two gene expression profiles showing specific patterns for low radiation doses 24 and 48 h after exposure. Since one-way ANOVA did not identify significantly regulated genes 24 h after low dose exposure, the first expression profile was constructed based on 24 genes that have at least a log_2 fold-change of 1 and a variance smaller than the median variance across all low dose samples measured 24 h after exposure. The second profile consists of 144 genes significantly regulated 48 h after exposure, as detected by one-way ANOVA. As shown in Figure 4.5 on the following page hierarchical clustering of both profiles show that at 24 and 48 h after exposure, all non-irradiated samples can be clearly discriminated from the irradiated samples. At 48 h after irradiation, the two radiation doses, 0.02 and 0.1 Gy, are correctly separated from each other as well. Twenty-four hours after irradiation, both radiation doses lead to very similar expression patterns such that one sample irradiated with 0.02 Gy is wrongly associated with the 0.1 Gy samples. Nevertheless, both heatmaps indicate that expression signatures appropriate for predicting low radiation doses might exist.

4.3.1 Biomarker signatures and prediction performances

One aim of this thesis is the identification of expression signatures for discriminating low radiation doses and to evaluate their prediction performance. To this end, I first identified 600 lists of significant radiation-responsive genes with the p-value-driven gene selection. Since the results for the medium and high dose ranges clearly show, that the consecutively applied fold-change-driven gene selection significantly improves the accuracy of radiation dose prediction (see Table 4.1 on page 56), I subsequently selected genes that display the highest fold-change between 0.02 and 0.1 Gy. Across all 600 lists obtained by the p-value-driven and consecutively applied fold-change-driven gene selection, 101 different genes were identified (*i.e.* union of all 600 lists, each comprising 20 genes with highest fold-change). Thus, the lists of radiation-responsive genes with highest fold-change consist of a varying composition of the 101 genes (see Supplementary Table A.1 on page 92).

The cross-validated performance evaluation revealed that the prediction of low radiation doses yielded an average sensitivity of 85.9%, an average specificity of 92.9% and an average overall success of 90.6%. A closer look at the prediction performances obtained for the single radiation doses indicates that the classification of samples irradiated with

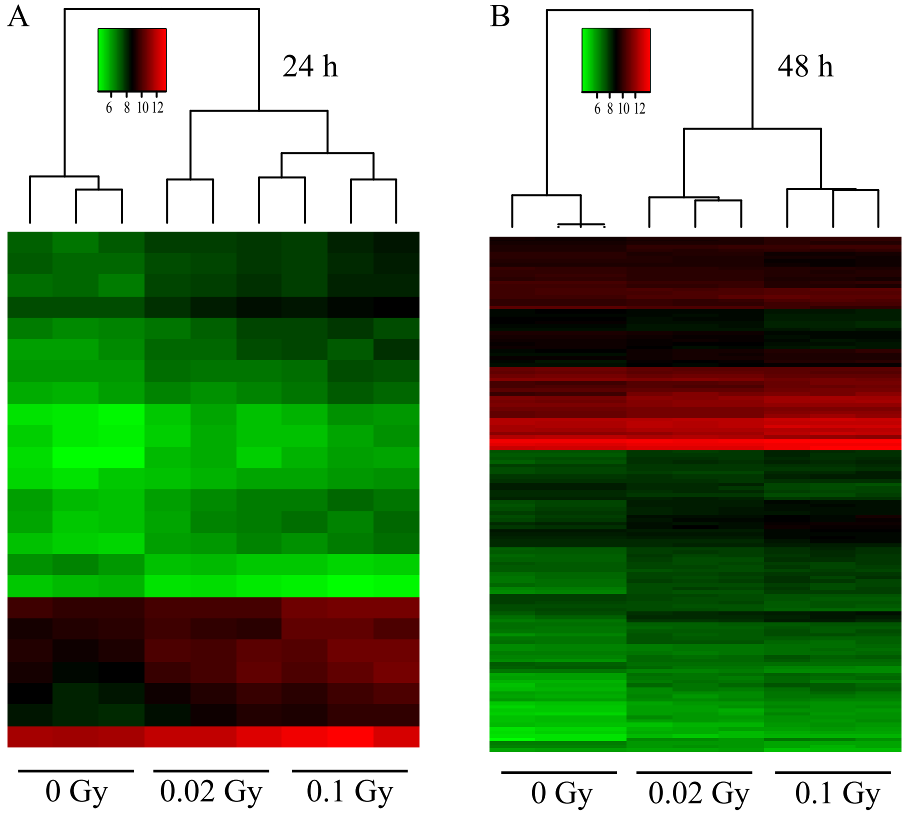


Figure 4.5: Heatmaps illustrating expression levels of radiation-induced of genes after low dose exposure. Genes (rows) showing an expression change after low dose exposure were selected to cluster samples (columns) irradiated with low radiation doses (0, 0.02, 0.1 Gy). High expression is depicted as red and low expression is depicted as green. Both heatmaps illustrate that all three radiation doses can be discriminated at both investigated time points after irradiation. **A:** Hierarchical clustering of DNA-microarray samples based on the expression levels of 24 genes with a $\log_2$ fold-change > 1 and a variance smaller than the median variance across all samples 24 h after low dose exposure (see also Supplementary Table A.2 on page 93). **B:** Hierarchical clustering of the DNA-microarray samples based on the expression levels of 144 significantly altered genes 48 h after low dose exposure (see Supplementary Table A.3 on page 99).

0.02 Gy is most problematic. Whereas 98.7% of all non-irradiated samples and 92.3% of all samples irradiated with 0.1 Gy are correctly predicted, only 66.7% of samples irradiated with 0.02 Gy are correctly assigned to their radiation dose (see Table 4.3).
By constructing the ensemble-based consensus signature I solved the difficulty of predicting the 0.02 Gy samples. An increased sensitivity of 95.6% (+ 9.7% compared to the performance of the *p*-value-driven and consecutively applied fold-change-driven gene selection), an increased specificity of 97.8% (+ 4.9%) and an increased overall success of 97% (+ 6.4%) was obtained. Remarkably, 100% of the non-irradiated and 0.1 Gy irradiated samples and as many as 86.7% of the 0.02 Gy irradiated samples are correctly predicted. The differences between the measured prediction performances based on the *p*-value-driven and consecutively applied fold-change-driven gene selection and the ensemble-based consensus signature are depicted in Table 4.3 on the next page. Moreover Table 4.3 demonstrates that the consensus signature leads to a significant improvement of the predictive power of low dose classification.
The consensus signature contains nine genes that are present in at least 500 of the 600 lists, each comprising 20 genes with highest fold-change between 0.02 and 0.1 Gy. Four of the nine genes are part of all 600 lists (see Table 4.4 on the following page). Interestingly, the well-known radiation-responsive gene *FDXR* is among these. As already mentioned before (see Section 4.2.2 on page 56), the function of FDXR protein is electron transfer from NADPH to cytochrome P450 via ferredoxin in mitochondria, and FDXR can be induced by DNA damage in a p53-dependent manner that sensitizes cells to ROS-mediated apoptosis (Liu and Chen, 2002). The genes *PFKFB3* and *LY6GG5C* are also included in the low dose consensus signature. *PFKFB3* is a key regulator of glycolysis (Okar and Lange, 1999), and *LY6G5C* belongs to a cluster of leukocyte antigen-6 (LY6) genes that are located in the major histocompatibility complex (MHC) class III region on chromosome 6 (Mallya et al., 2002). The function of the other genes identified is as yet unknown.

4.3.2 Confirming our results by qRT-PCR analysis

With my DNA-microarray analysis, I identified nine genes that are suitable for radiation dose prediction 24 and 48 h after low dose exposure with DNA-microarray analysis (see Table 4.4 on the following page). To validate the gene expression alterations, we selected three of the nine candidate biomarker genes, namely *FDXR*, *PFKFB3* and *LY6G5C*, for qRT-PCR analysis. The qRT-PCR analysis was based on non-pooled RNA samples of four healthy donors.
As shown in Figure 4.6 on page 65 an increasing gene expression was measured for *FDXR* 24 h after irradiation with 0.02 and 0.1 Gy, and 48 h after irradiation with 0.1 Gy in comparison to the non-irradiated control. In contrast, a downregulation of

Method	Dose (Gy)	Sensitivity	Specificity	Overall Success
Fold-change-driven				
	0	98.7	97	97.6
	0.02	66.7	95.5	85.9
	0.1	92.3	86.3	88.3
	Avg	85.9	92.2	90.6
Consensus signature				
	0	100	100	100
	0.02	86.7	100	95.6
	0.1	100	93.3	95.6
	Avg	95.6	97.8	97.0

Table 4.3: Prediction performances obtained for the low radiation doses. The performances yielded by the *p*-value-driven and consecutively applied fold-change-driven gene selection and the ensemble-based consensus signature are compared. The sensitivities, specificities and overall successes obtained for each radiation dose (0, 0.02, 0.1 Gy) and averaged over the complete dose range (Avg) are given in percent.

Agilent ID	Gene Symbol	Frequency of Selection
A_24_P506680	N/A	600
A_24_P375205	*MKL2*	600
A_23_P38154	*FDXR*	600
A_24_P332081	*JAKMIP3*	600
A_24_P111096	*PFKFB3*	599
A_32_P138939	N/A	581
A_23_P419868	*FLJ35379*	544
A_32_P20997	N/A	541
A_24_P7584	*LY6G5C*	532

Table 4.4: Ensemble-based consensus signature for predicting low radiation doses. All genes included in the ensemble-based consensus signature constructed for the low dose range (0, 0.02, 0.1 Gy) are listed. The frequency of selection indicates how often a gene was among the 20 selected genes with highest fold-change between 0.02 and 0.1 Gy within 100 repeated 6-fold-cross-validations. This Table is adapted from Knops et al. (2012).

FDXR is induced 48 h after irradiation with 0.02 Gy. *PFKFB3* reveals a downregulation at all examined time points and radiation doses that continued to decline with increasing radiation doses 24 h after exposure. Very similar gene expression values were detected 24 h after irradiation with 0.1 Gy and 48 h after irradiation with 0.02 and 0.1 Gy. However, *LY6G5C* exhibits a detectable (but not statistically significant) upregulation 24 h after irradiation with 0.02 Gy.

In summary, we successfully validated the gene expression alterations detected by microarray analysis with our qRT-PCR analysis based on non-pooled RNA samples from four healthy donors. Moreover, the qRT-PCR analysis revealed relatively low inter-individual gene expression variations in examined genes among the four donors, especially for *FDXR* 24 h and for *PFKFB3* 48 h after irradiation with 0.1 Gy.

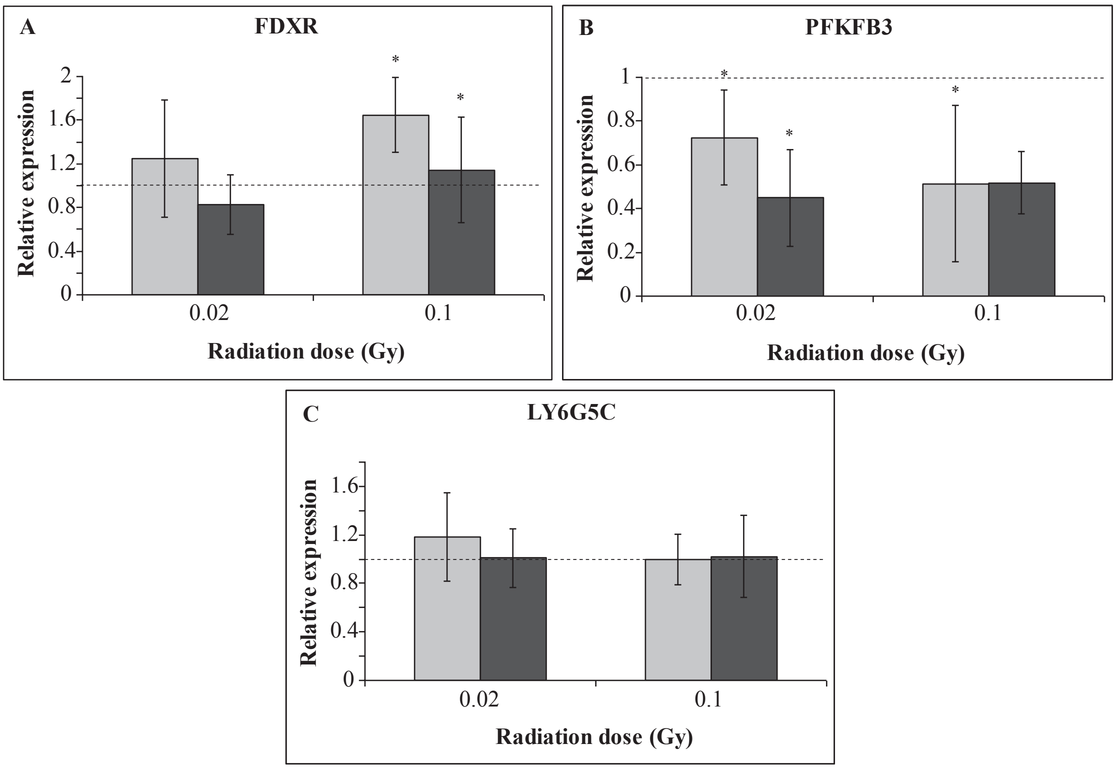


Figure 4.6: Illustration of the mean value of relative expression of six radiation-responsive genes in non-pooled, individual RNA samples based on qRT-PCR after low dose exposure. The relative expressions 24 h (light grey bars; $n = 4$) and 48 h (dark grey bars; $n = 4$) after low dose exposure (0.02, 0.1 Gy) measured by qRT-PCR based on the expression of the non-irradiated control (dashed line) are shown. A significant rise of the gene expression from one dose to the next higher dose is marked by an asterisk (p-value < 0.05). **A:** *FDXR* reveals the strongest upregulation 24 h after irradiation that increases with increasing dose. **B:** For *PFKFB3* at all examined time points and radiation doses a downregulation is detected. **C:** *LY6G5C* shows a detectable upregulation 24 h after irradiation with 0.02 Gy. Error bars indicate the standard error of the mean for $n = 4$ independent experiments.

4.4 Comparison to other gene expression-based biodosimetric studies

In recent years, many studies investigated radiation-induced expression changes as a tool for biodosimetric applications (Brengues et al., 2010; Dressman et al., 2007; Gruel et al., 2006; Meadows et al., 2008; Paul and Amundson, 2008). However, a direct comparison to the results of my work is intricate due to varying experimental settings such as different cell types, radiation doses or time points post irradiation. Additionally, the technology used for transcriptional profiling (*e.g.* microarrays or qRT-PCR) or the utilized microarray platform (*e.g.* Agilent or Affymetrix) differs across the studies. Different statistical and bioinformatics-driven analysis have been pursued depending on the purpose and the main focus of the studies. In the following, I consider our candidate gene signatures for radiation dose prediction in the context of the results presented by similar gene expression-based biodosimetric studies. I separately discuss the here identified signatures (1) for the medium to high radiation doses (0.5-4 Gy) and (2) for the low radiation doses (0.02-0.1 Gy).

1. Medium to high radiation doses: A very similar study with respect to the experimental setting and statistical analysis was conducted by Paul and Amundson (2008). The authors profiled the gene expression of γ-irradiated human PBLs from ten healthy donors using Agilent Whole Human Genome Oligo Microarrays (G4112A). They monitored the transcriptional level of genes at both 6 and 24 h after exposure to doses of 0, 0.5, 2, 5, and 8 Gy. Based on the experimental data they extracted a set of 74 radiation-induced genes with which four dose ranges (*i.e.* non-irradiated, 0.5, 2 and 5–8 Gy) could be predicted with a sensitivity of 98%. With my computational framework for biomarker discovery and radiation dose prediction, I extracted 16 potential biomarker genes (see Table 4.2 on page 57), with which five radiation doses (*i.e.* non-irradiated, 0.5, 1, 2 and 4 Gy) can be predicted at three time points (*i.e.* 12, 24 and 48 h) with a sensitivity of 95.7% (see Table 4.1 on page 56). Five of the 16 genes, namely *VWCE*, *TNFSF4* *FDXR*, *ISGL20L1* and *C8orf38*, are also part of the 74 gene signature identified by Paul and Amundson (2008). Three additional genes of our 16 biomarker candidates were reported to be differentially expressed by Paul and Amundson (2008), but were not included in their 74-gene signature: *PHLDA3* and *RP4-742C19.3* showed significant gene expression changes 6 h and *TCL1A* 6 h and 24 h after irradiation. Note, that therewith also three genes from our consensus signature, identified for the medium to high dose range (see Section 4.2.2 on page 56), are part of their 74-gene signature (*i.e.* *VWCE*, *TNFSF4* and *FDXR*).

Several of our candidate biomarker genes are also reported by other biodosimetric studies, even though not all explicitly utilized them for estimating the radiation dose. For example, *FDXR* was shown to be radiation-induced after exposure to low LET radiation in CD4$^+$ T-lymphocytes (Mori et al., 2005), peripheral blood samples (Budworth et al., 2012; Manning et al., 2013), primary human fibroblast cell lines (Kis et al., 2006), non-immortalized human T cells (Pogosova-Agadjanyan et al., 2011) and dividing lymphocytes as well as peripheral blood leukocytes (Kabacik et al., 2011). Besides an upregulation of *FDXR*, *TNFSF4*, *PHLDA3* and *ISGL201* were also shown to be responsive to radiation in previous publications (Mori et al., 2005; Pogosova-Agadjanyan et al., 2011). In summary, our results show an evident overlap with existing investigations. However, we also extracted genes, such as *SPATA18*, which have not previously been used for estimating medium to high radiation doses.

2. Low radiation doses: In recent years, the interest in elucidating the effects of low dose exposure on the cellular transcription has considerably grown (Fachin et al., 2007, 2009; Jin et al., 2008; Pogosova-Agadjanyan et al., 2011; Wyrobek et al., 2011). As described in Section 4.3 on page 61, I identified potential biomarker signatures with which low radiation doses (0, 0.02, 0.1 Gy) can be accurately predicted.
For the low dose range, *FDXR* is the only gene of our signature that was also detected by other researchers analyzing gene expression changes after irradiation. *FDXR* is part of (i) the gene signature, which I extracted for the medium to high dose range (see also previous paragraph) and additionally of (ii) the 74-gene signature identified by Paul and Amundson (2008). Interestingly, *FDXR* is also one of nine biomarkers which were proposed only recently by Manning et al. (2013) for estimating low radiation doses (0.05–1 Gy). The authors investigated the response of mostly p53-regulated genes to X-ray exposure. They measured their gene expression in *ex vivo* irradiated blood from healthy donors by Multiplex qRT-PCR and established and characterized their exposure-response relationships. By using different combinations of nine biomarker candidates they showed that the gene set of *FDXR*, *DDB2* and *CCNG1* gave the best estimate for low radiation doses.
Except *FDXR*, none of our signature genes was previously reported in the context of radiation. Nevertheless, many studies provided evidence that ionizing radiation leads to significant transcriptional changes even after low dose exposure. It is for this reason that gene expression signatures, such as the one proposed in the present work, are a promising tool also for predicting low radiation doses.

4.5 Summary and conclusion

In this chapter, an ensemble-based strategy for the identification of stable gene expression signatures functioning as radiation biodosimeters is described. The development and implementation of a computational framework for biomarker discovery and radiation dose prediction has led me to more carefully consider the problem of reproducibility of gene expression-based biomarkers. For many years a good prediction accuracy was the most important decision criterion in selecting promising biomarker candidates. Only the last few years, researcher stressed that the ascertainment of a true prediction outcome is no less important than the reproducibility of the results (Taylor et al., 2008). To this end, I proposed a principle approach for selecting biomarker candidates from high-throughput gene expression data, whose algorithmic design is intended to support reproducibility in terms of selecting genes which are stable with respect to variations within experimental data. A fundamental component of my approach is the repeated cross-validation procedure which generates different sample subsets of the underlying data, functioning as training sets and test sets in the subsequent procedure. First, I selected for each training set a subset of radiation-induced genes with high changes in gene expression after radiation exposure. It has been previously shown by Fan et al. (2010) that this type of feature selection, which I here refer to as *p*-value-driven and fold-change-driven gene selection, is known to support the stability of selected genes. Second, I aggregated all sets of biomarker candidates selected in the first step, which are likely to differ from each other because they originate from slightly varying sample subsets, to one ensemble output. The ensemble-based consensus signature contains genes which were most frequently selected by the *p*-value-driven and fold-change-driven gene selection across all training sets. The advantage of this approach is that noise is eliminated because only genes are selected which are consistently found to be significant (Davis et al., 2006). Furthermore, they are presumed to be highly stable in terms of data variations by subsampling (Abeel et al., 2010; He and Yu, 2010) and most relevant to sample prediction (Baek et al., 2009; Chen et al., 2007).

An important aspect of our microarray experiment with regard to its application to radiation biodosimetry is our approach of pooling the blood samples from six donors. Pooling mRNA of a group of donors is one option to control the costs of a DNA-microarray experiment and can be useful or even necessary when there is not enough isolated mRNA from each donor to hybridize individual DNA-microarrays (Peng et al., 2003). For us the most important reason was a different one. We deliberately opted for pooling the blood samples in order to identify radiation-induced expression profiles which are independent from individual patterns and outliers. This strategy of reducing the effect of biological variation (Kendziorski et al., 2003) and thus to identify gene patterns generalizable to estimate the radiation dose of individuals was also applied in previously published biodosimetric studies (Kis et al., 2006; Port et al., 2007).

My study aims at the identification of preferably small and stable gene signatures allowing an accurate prediction of low (0, 0.02, 0.1 Gy) and medium to high radiation doses (0, 0.5, 1, 2, 4 Gy). To this end, I applied my computational framework I developed for biomarker discovery and radiation dose prediction to our microarray gene expression data.

For the medium to high radiation doses, my bioinformatics-driven analysis revealed the following key findings: with the above mentioned p-value-driven and fold-change-driven gene selection, I identified 16 biomarker candidates covering the presumed dose range (0.5–4 Gy) and time frame (6–48 h) of large scale radiation accidents. Using the 16 genes for radiation dose prediction, an averaged sensitivity of 95.7% was reached. With the consecutively built ensemble-based consensus signature, containing seven of the 16 biomarker candidates, a decreased averaged sensitivity of 89.1% was obtained. Since stability of biomarkers is an important issue, as it may improve the success of subsequent validations on external datasets (Abeel et al., 2010), one may have to accept a trade-off between improved reproducibility against a decreased prediction accuracy.

One of the seven genes included in the consensus signature, namely *FDXR*, is amongst the most commonly identified radiation-responsive genes in the literature (Oh et al., 2012); the six remaining genes were less frequently (*TNFSF4*, *PHLDA3*, *VWCE*) or not reported before in the context of radiation biodosimetry (*DOK7*, *SPATA18*, *LGR6*). The detailed function of these genes as well as their implication in radiation response is still not fully resolved.

Another objective of my analysis is to investigate whether my computational framework for biomarker discovery and radiation dose prediction is a valid tool for identifying promising biomarker candidates also for the low dose range (0, 0.02, 0.1 Gy). Several studies previously reported that low dose exposure leads to significant changes in gene expression, but whether gene signatures can be utilized for predicting low radiation doses is an open question and not sufficiently investigated so far.

The results of my analysis support the idea that samples irradiated with low radiation doses can be discriminated based on gene expression levels. Using radiation-responsive genes which display high gene expression changes between 0.02 and 0.1 Gy for radiation dose prediction I obtained an averaged sensitivity of 85.9%. In contrast to the results obtained for the medium to high dose range, the ensemble-based consensus signature yielded an improved prediction performance for the low dose range. The consensus signature, consisting of nine genes, enabled a radiation dose prediction with an averaged sensitivity of 95.6% and revealed that low dose irradiation already leads to clear gene expression alterations compared to the non-irradiated control samples. *FDXR* was the only gene of the ensemble-based consensus signature for radiation dose prediction after low dose exposure that was also detected in several other studies. The function of the remaining is as yet unknown.

As already mentioned above, I identified gene signatures for radiation dose prediction based on microarray data of pooled and irradiated blood samples. Of course, it is indispensable for future biodosimetric applications that the discovered gene expression-based radiation biodosimeters allow an estimation of individual radiation doses. Thus, we additionally analyzed the gene expression changes with qRT-PCR analysis based on non-pooled blood samples and successfully validated most of our discussed genes.

A comparison of the potential biomarker signatures identified for the low dose range and for the medium to high dose range revealed the following interesting finding: for the medium to high dose range, a higher percentage of genes are present in all lists of radiation-responsive genes with highest maximal fold-change than for the low dose range. Seven of the 16 genes identified by the p-value-driven and consecutively applied fold-change-driven gene selection are part of all lists identified for the medium to high radiation doses, whereas only four of the 101 genes are part of all lists for the low radiation doses. Since the ratio of the intersection of gene lists to their union is a common similarity measure to assess the stability of feature selection results (He and Yu, 2010), it can be concluded that the potential biomarker signature for the medium to high dose range is characterized by a higher stability with respect to sampling variations. The lower stability observed for the results for low radiation doses could have two causes: first, we used a smaller number of DNA-microarray samples for the classification of low radiation doses than for classification of the medium and high radiation doses. As stated by He and Yu (2010) the relatively small number of samples in high-dimensional data is one of the main sources of the instability problem in feature selection and it has been previously verified that the overlap between independently developed gene signatures was increased linearly with more samples (Kim, 2009). Second, as shown in Chapter 3 on page 27, low radiation doses result in less pronounced gene expression changes. This is why there is a higher chance that noise or natural variations in gene expression superimpose transcriptional modulations induced by ionizing radiation. Hence, my gene selection procedure based on the statistical analysis of expression levels may not observe all changes in the sample subsets.

My computational framework for biomarker discovery and radiation dose prediction incorporates a frequency-based selection of the top-ranked genes over different lists obtained by cross-validation. As already described by Boulesteix and Slawski (2009) this approach heavily depends on the arbitrarily fixed threshold k, defining the number of genes at the top of each list used to construct the ensemble-based consensus signatures. To overcome the problem of setting a fixed threshold, one could successively try several thresholds and choose one cutoff resulting in both a high prediction accuracy and a high gene signature stability. Note that the optimization of model parameters in classification processes, like choosing the best threshold k, necessitates the implementation of a nested cross-validation scheme (Varma and Simon, 2006). The inner loop of a nested cross-validation is used to perform the parameter tuning while the outer loop is used to evaluate the performance of the classification. Since our ensemble-based consensus signatures meet the requirements of an accurate dose prediction and a preferably high stability, I decided not to optimize the threshold k by incorporating a nested cross-validation, which would have significantly increased the complexity of my

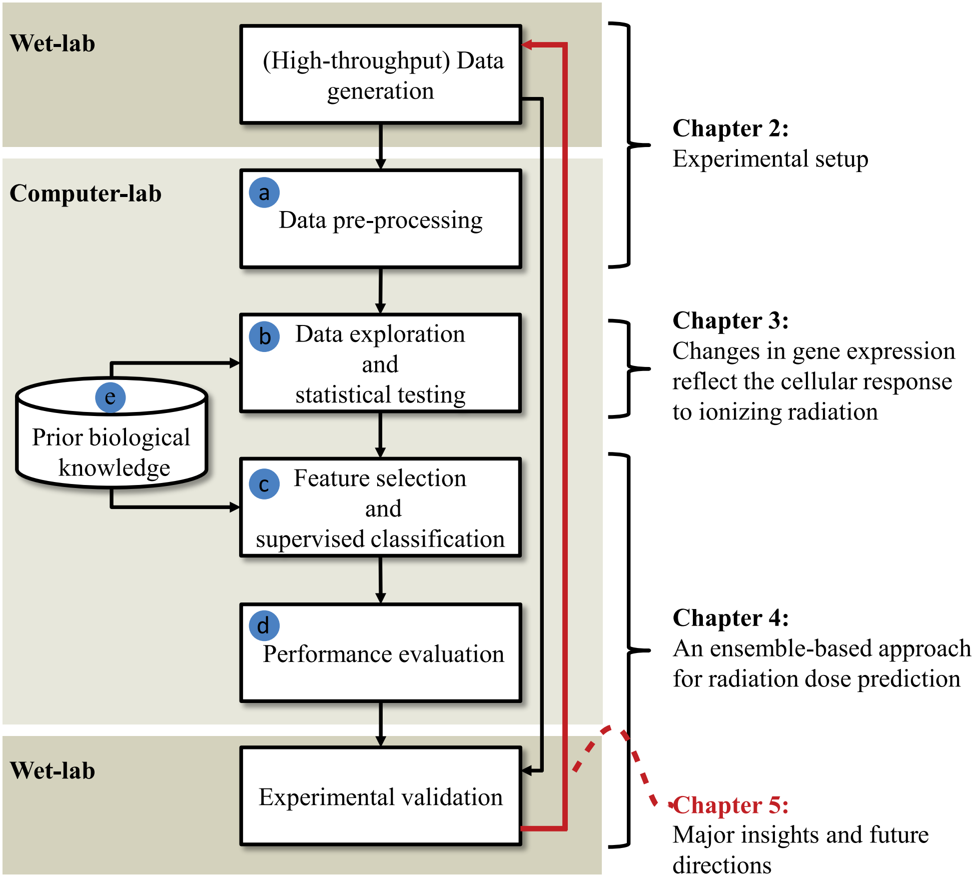
computational framework for biomarker discovery and radiation dose prediction.
In summary, the genes identified for the medium to high dose range are potential biomarkers, which are particularly suitable for dose level discrimination in a time frame that would be appropriate for life-saving medical triage. Even though the discrimination of low radiation doses is not essential for acute medical decision-making, a biodosimetry method based on gene expression profiles to determine even low radiation doses would be beneficial to assess the long-term health risks for a large number of exposed individuals. By applying my framework for biomarker discovery and radiation dose prediction to our DNA-microarray samples, measuring the gene expression in human PBLs after low dose exposure, I identified a potential biomarker signature with which low radiation doses could be accurately assessed.

Chapter 5

Major insights and future directions

Synopsis

In this chapter, the major findings of this thesis are summarized with respect to the workflow I passed through in the course of my dissertation (see below). Furthermore, directions for improvements of the here presented work and implications for future research are discussed. In particular, inspired by the obtained results, new wet-lab experiments are proposed and their impact for future biodosimetry studies are considered.


Wet-lab
(High-throughput) Data generation
Computer-lab
a
Data pre-processing
b
Data exploration and statistical testing
e
Prior biological knowledge
c
Feature selection and supervised classification
d
Performance evaluation
Wet-lab
Experimental validation
Chapter 2:
Experimental setup
Chapter 3:
Changes in gene expression reflect the cellular response to ionizing radiation
Chapter 4:
An ensemble-based approach for radiation dose prediction
Chapter 5:
Major insights and future directions

In the introduction of this thesis I presented a general workflow for high-throughput data analysis and biomarker discovery (see Figure 1.1 on page 4). With the aim to identify gene signatures functioning as radiation biodosimeters, I passed through all steps of this workflow by combining the skills and resources of different disciplines, including molecular biology, computer science and bioinformatics. The individual steps, their corresponding objectives and major biological findings, which were discussed in detail in the course of this thesis, can be summarized as follows:

1. **High-throughput data generation:** With whole genome DNA-microarrays we monitored gene expression levels in human PBLs after *ex vivo* γ-irradiation. This data provide the basis for my bioinformatics-driven analysis pipeline and thus for the biological findings presented in this thesis.

2. **Statistical and functional data analysis:** With my systematic and comparative DNA-microarray analysis, I investigated radiation-induced transcriptional changes in order to characterize the DNA damage response to low, medium and high dose exposure. From the analysis, we can see that radiation dose and time after exposure substantially influence the number of radiation-induced genes as well as the affected pathways and molecular mechanisms. Furthermore, the results indicate that even acute low dose exposure causes a well-defined physiological responses in human PBLs.

3. **Bioinformatics-driven discovery of biomarker signatures:** I developed a combined statistical and bioinformatics-driven framework in order to identify radiation-responsive genes in human lymphocytes as a tool for radiation biodosimetry. By implementing a cross-validation- and frequency-based gene selection procedure, I constructed two ensemble-based consensus signatures for radiation dose prediction. The genes of both signatures show a high stability against small variations in our data. From the performance evaluation of my supervised classification, I conclude that the two consensus signatures allow an accurate retrospective estimation of low and medium to high radiation doses up to 48 h after exposure.

The results of my thesis clearly show that the dose-response relationship of selected genes allow a precise *in vitro* estimation of radiation doses. But what impact or relevance do our findings have for practical radiation biodosimetry?

In the present work, it has been demonstrated that gene expression signatures allow a discrimination of radiation doses as low as 0.02 Gy. Being able to determine the radiation dose after low dose exposure is an important task for public health science (Jin et al., 2008). While the annual dose limit for persons occupationally exposed to ionizing radiation is 20 mSv (Sudprasert et al., 2006), one single computer tomography scan results in radiation absorbed doses of 12 mGy (body) to 20 mGy (head) (Huda and Ogden, 2008). Additionally, the majority of affected individuals in a large-scale nuclear accident will probably receive total-body doses significantly smaller than 100 mSv. In both cases, determining the radiation dose is critical for epidemiological surveys of health effects. Together with the results we derived for the medium to high radiation doses we showed that gene expression-based biodosimetry has the potential to cover a large range of radiation doses (0.02-4 Gy) at a window of time essential for medical decision making after a radiation accident.

Besides a well-defined dose-response relationship, a low minimal detectable dose and the coverage of a large dose range, a radiation biodosimeter should be specific to ionizing radiation and not be affected by confounding factors (Romm et al., 2009; Simon et al., 2010; Voisin et al., 2004). Both issues were not addressed in the present thesis, but only recently discussed in related studies:

Knops (2013) compared the effects on gene expression caused by two DNA damage-inducing agents, namely 4-Acetamidophenol and Mitomycin C, with the effects caused by ionizing radiation. Although the treatment with 4-Acetamidophenol and Mitomycin C led to an increased expression of three genes contained in the here presented ensemble-based consensus signature for predicting medium to high radiation doses (*i.e. FDXR*, *PHLDA3* and *TNFSF4*), the comparison revealed that ionizing radiation causes specific expression patterns which can be distinguished from those induced by 4-Acetamidophenol or Mitomycin C. Knops (2013) therefore concluded that the cellular mechanisms induced by chemotherapeutic agents and those induced by radiation exposure are independent from each other.

In a recent study, Paul and Amundson (2011) investigated the potential impact of smoking on the gene expression response to radiation exposure. The authors evaluated the ability of a previously identified 74-gene expression signature (Paul and Amundson, 2008) to predict the radiation exposure level of *ex vivo* irradiated blood samples of smokers and non-smokers of both genders. Their results showed that the accuracy of their previously defined 74-gene signature in radiation dose prediction was unaffected by differences in smoking status or gender. Please note that five of the nine genes contained in our ensemble-based consensus signature are also included in their expression signature.

The translation of the *in vitro* results presented in this dissertation to *in vivo* is an indispensable step to transfer our findings into a clinical application like a diagnostic devise for biodosimetry. Most of the published gene expression-based biodosimetry studies refer to *ex vivo* results and did not investigate their comparability to *in vivo*. However, Paul et al. (2011) tested the ability of gene expression signatures to predict radiation doses received *in vivo* in a population of patients undergoing total body irradiation. Since their previously identified *ex vivo* signature (Paul and Amundson, 2008) predicted *in vivo* exposure with 100% accuracy and dose level with up to 98% accuracy, the authors concluded that using *ex vivo* studies, like the one presented in this thesis, are a promising and suitable approach for the discovery of gene signatures applicable in future biodosimetry applications.

At this point my work inspires new wet-lab experiments, illustrated by the iterative cycle integrated in the workflow for data-driven biomarker discovery presented in Figure 1.1 on page 4. Customized DNA-microarrays, measuring the expression levels of those genes which are part of our two ensemble-based consensus signatures, could be used to perform an *in vivo* confirmation of our results. Hereby, the peripheral blood of patients having a computer tomography scan or undergoing total body irradiation in preparation for stem cell transplantations would provide suitable experimental data for an *in vivo* validation of our gene signatures for the low and medium to high radiation doses respectively (Amundson et al., 2004).

Once the *in vivo* applicability of a gene signature is successfully confirmed, the genes could be utilized for a biodosimetry device, like the one proposed by Brengues et al. (2010), which measures the expression levels of 14 pre-selected genes. The authors state that only a fingerstick of blood is needed for screening and the data can be delivered in less than 12 h. Hence, such a gene expression-based biodosimetry device combines several characteristics, like minimally invasive sampling, low sample costs and a standardized, automated data processing, which are favorable or even required for an effective and fast screening of potentially exposed individuals after a large radiation accident.

The applicability of gene expression-based radiation dose prediction to practical biodosimetry largely depends on the reproducibility of the utilized gene signatures. The design of my computational framework for biomarker discovery and radiation dose prediction intends to support biomarker stability and therewith reproducibility. At this point, I would like to emphasize that the here presented strategy is considered to be one possibility to enhance biomarker stability (Boulesteix and Slawski, 2009; He and Yu, 2010), though not of course the only one. For example, one promising idea to improve the stability of biomarker signatures is to incorporate prior biological knowledge (*e.g.* functional annotations, protein-protein interactions and expression correlation among genes) into the process of feature selection (Haury et al., 2010; Sanavia et al., 2012). As we have seen in the course of this thesis, the process for the identification of potential biomarkers comprises several steps (see Figure 1.1 on page 4), whereas each step can influence the stability of the signatures identified. Studies which investigated the impact of different approaches on the stability, often focussed on only one of these steps (like feature selection, classification or the integration of prior knowledge) and did not consider possible dependencies between them (Davis et al., 2006; Fan et al., 2010; Haury et al., 2011; Sanavia et al., 2012). To provide a more complete picture of this issue, more research is needed to identify and systematically evaluate factors influencing biomarker stability.

In conclusion, although the gene expression-based prediction of radiation doses is a young branch of radiation biodosimetry, it holds great promise for future applications. With my combined statistical and bioinformatics-driven framework for biomarker discovery and radiation dose prediction, I identified candidate biomarker signatures, which lay the foundation for the development of advanced biodosimetry devices. The latter combines characteristics which are of particular value in situations where a large number of exposed people have to be screened in order to guide prompt medical decision-making or to assess long-term health risks of radiation exposure. The present work demonstrates that radiation-responsive genes in human lymphocytes are a promising biodosimetric tool, even though it will need many years of joint research before gene expression-based biodosimetry becomes a standardized, validated method.

Bibliography

T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26 (3):392–398, 2010. (cited on pages 51, 68 and 69)

D. W. Aha, D. Kibler, and M. K. Albert. Instance-based learning algorithms. *Machine learning*, 6(1):37–66, 1991. (cited on page 49)

E. A. Ainsbury, E. Bakhanova, J. F. Barquinero, M. Brai, V. Chumak, V. Correcher, F. Darroudi, P. Fattibene, G. Gruel, I. Guclu, S. Horn, A. Jaworska, U. Kulka, C. Lindholm, D. Lloyd, A. Longo, M. Marrale, O. Monteiro Gil, U. Oestreicher, J. Pajic, B. Rakic, H. Romm, F. Trompier, I. Veronese, P. Voisin, A. Vral, C. A. Whitehouse, A. Wieser, C. Woda, A. Wojcik, and K. Rothkamm. Review of retrospective dosimetry techniques for external ionising radiation exposures. *Radiation Protection Dosimetry*, 147(4):573–592, 2011. (cited on page 14)

F. Al-Shahrour, R. Díaz-Uriarte, and J. Dopazo. Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*, 20(4):578–580, 2004. (cited on page 32)

S. A. Amundson, M. Bittner, Y. Chen, J. Trent, P. Meltzer, and A. J. Fornace. Fluorescent cDNA microarray hybridization reveals complexity and heterogeneity of cellular genotoxic stress responses. *Oncogene*, 18(24):3666–3672, 1999a. (cited on page 16)

S. A. Amundson, K. T. Do, and A. J. Fornace. Induction of stress genes by low doses of gamma rays. *Radiation Research*, 152:225–231, 1999b. (cited on page 16)

S. A. Amundson, K. T. Do, S. Shahab, M. Bittner, P. Meltzer, J. Trent, and A. J. Fornace. Identification of potential mRNA biomarkers in peripheral blood lymphocytes for human exposure to ionizing radiation. *Radiation Research*, 145(3):342–346, 2000. (cited on page 16)

S. A. Amundson, M. B. Grace, C. B. McLeland, M. W. Epperly, A. Yeager, Q. Zhan, J. S. Greenberger, and A. J. Fornace. Human in vivo radiation-induced biomarkers: gene expression changes in radiotherapy patients. *Cancer Research*, 64(18):6368–6371, 2004. (cited on page 76)

S. A. Amundson, K. T. Do, L. C. Vinikoor, R. A. Lee, C. A. Koch-Paiz, J. Ahn, M. Reimers, Y. Chen, D. A. Scudiero, J. N. Weinstein, J. M. Trent, M. L. Bittner, P. S. Meltzer, and A. J. Fornace. Integrating global gene expression and radiation survival parameters across the 60 cell lines of the National Cancer Institute Anticancer Drug Screen. *Cancer Research*, 68(2):415–424, 2008. (cited on page 58)

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29, 2000. (cited on pages 7 and 32)

A. J. Atkinson, W. A. Colburn, V. G. DeGruttola, D. L. DeMets, G. J. Downing, D. F. Hoth, J. A. Oates, C. C. Peck, R. T. Schooley, B. A. Spilker, J. Woodcock, and S. Zeger. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology and Therapeutics*, 69(3):89–95, 2001. (cited on page 3)

S. Baek, C. A. Tsai, and J. J. Chen. Development of biomarker classifiers from high-dimensional data. *Briefings in Bioinformatics*, 10(5):537–546, 2009. (cited on pages 45, 51 and 68)

A. Bajinskis. *Studies of DNA repair strategies in response to complex DNA damages*. PhD thesis, Stockholm University, Department of Genetics, Microbiology and Toxicology, 2012. (cited on page 28)

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. (cited on pages 31 and 52)

S. Boldt, K. Knops, R. Kriehuber, and O. Wolkenhauer. A frequency-based gene selection method to identify robust biomarkers for radiation dose prediction. *International Journal of Radiation Biology*, 88(3):267–276, 2012. (cited on page 56)

C. Bolognesi, C. Balia, P. Roggieri, F. Cardinale, P. Bruzzi, F. Sorcinelli, F. Lista, R. D'Amelio, and E. Righi. Micronucleus test for radiation biodosimetry in mass casualty events: evaluation of visual and automated scoring. *Radiation Measurements*, 46(2): 169–175, 2011. (cited on page 14)

C. Bornstein, R. Brosh, A. Molchadsky, S. Madar, I. Kogan-Sakin, I. Goldstein, D. Chakravarti, E. R. Flores, N. Goldfinger, R. Sarig, and V. Rotter. *SPATA18*, a spermatogenesis-associated gene, is a novel transcriptional target of p53 and p63. *Molecular and Cellular Biology*, 31(8):1679–1689, 2011. (cited on page 58)

A. L. Boulesteix and M. Slawski. Stability and aggregation of ranked gene lists. *Briefings in Bioinformatics*, 10(5):556–568, 2009. (cited on pages 44, 70 and 76)

M. Brengues, B. Paap, M. Bittner, S. Amundson, B. Seligmann, R. Korn, R. Lenigk, and F. Zenhausern. Biodosimetry on small blood volume using gene expression assay. *Health Physics*, 98(2):179–185, 2010. (cited on pages 16, 66 and 76)

D. J. Brenner, R. Doll, D. T. Goodhead, E. J. Hall, C. E. Land, J. B. Little, J. H. Lubin, D. L. Preston, R. J. Preston, J. S. Puskin, E. Ron, R. K. Sachs, J. M. Samet, R. B. Setlow, and M. Zaider. Cancer risks attributable to low doses of ionizing radiation: assessing what we really know. *Proceedings of the National Academy of Sciences of the United States of America*, 100(24):13761–13766, 2003. (cited on page 12)

H. Budworth, A. M. Snijders, F. Marchetti, B. Mannion, S. Bhatnagar, E. Kwoh, Y. Tan, S. X. Wang, W. F. Blakely, M. Coleman, L. Peterson, and A. J. Wyrobek. DNA repair and cell cycle biomarkers of radiation exposure and inflammation stress in human blood. *PLoS ONE*, 7(11):e48619, 2012. (cited on pages 36 and 67)

J. Cameron. Radiation dosimetry. *Environmental Health Perspectives*, 91:45, 1991. (cited on page 12)

H. G. Campbell, R. Mehta, A. A. Neumann, C. Rubio, M. Baird, T. L. Slatter, and A. W. Braithwaite. Activation of p53 following ionizing radiation, but not other stressors, is dependent on the proline-rich domain (PRD). *Oncogene*, 32(7):827–836, 2013. (cited on page 36)

J. J. Chen, C. A. Tsai, S. Tzeng, and C. H. Chen. Gene selection with multiple ordering criteria. *BMC Bioinformatics*, 8(1):74, 2007. (cited on pages 52 and 68)

X. Cui and G. A. Churchill. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology*, 4(4):210, 2003. (cited on page 6)

N. Dainiak, J. K. Waselenko, J. O. Armitage, T. J. MacVittie, and A. M. Farese. The hematologist and radiation casualties. *Hematology*, 2003(1):473–496, 2003. (cited on page 13)

C. A. Davis, F. Gerick, V. Hintermair, C. C. Friedel, K. Fundel, R. Kuffner, and R. Zimmer. Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics*, 22(19):2356–2363, 2006. (cited on pages 52, 68 and 76)

S. Derveaux, J. Vandesompele, and J. Hellemans. How to do successful gene expression analysis using real-time PCR. *Methods*, 50(4):227–230, 2010. (cited on page 23)

Y. Ding and D. Wilkins. The effect of normalization on microarray data analysis. *DNA and Cell Biology*, 23(10):635–642, 2004. (cited on page 24)

H. Dressman, G. Muramoto, N. Chao, S. Meadows, D. Marshall, G. Ginsburg, J. Nevins, and J. Chute. Gene expression signatures that predict radiation exposure in mice and humans. *PLoS Medicine*, 4(4):e106, 2007. (cited on pages 55 and 66)

R. Du, C. Huang, Q. Bi, Y. Zhai, L. Xia, J. Liu, S. Sun, and D. Fan. URG11 mediates hypoxia-induced epithelial-to-mesenchymal transition by modulation of E-cadherin and beta-catenin. *Biochemical and Biophysical Research Communications*, 391(1):135–141, 2010. (cited on page 58)

S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87, 2002. (cited on page 8)

S. Dudoit, J. P. Shaffer, and J. C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103, 2003. (cited on pages 6 and 7)

L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178, 2005. (cited on page 44)

L. Ein-Dor, O. Zuk, and E. Domany. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 103(15):5923–5928, 2006. (cited on pages 44 and 47)

A. R. Ennos. *Statistical and data handling skills in biology*. Pearson Education, 2007. (cited on pages 30 and 31)

W. Etienne, M. H. Meyer, J. Peppers, and R. A. Meyer. Comparison of mRNA gene expression by RT-PCR and DNA microarray. *Biotechniques*, 36(4):618–620, 2004. (cited on page 59)

A. L. Fachin, S. S. Mello, P. Sandrin-Garcia, C. M. Junta, E. A. Donadi, G. A. Passos, and E. T. Sakamoto-Hojo. Gene expression profiles in human lymphocytes irradiated in vitro with low doses of gamma rays. *Radiation Research*, 168(6):650–665, 2007. (cited on pages 34 and 67)

A. L. Fachin, S. S. Mello, P. Sandrin-Garcia, C. M. Junta, T. Ghilardi-Netto, E. A. Donadi, G. A. da Silva Passos, and E. T. Sakamoto-Hojo. Gene expression profiles in radiation workers occupationally exposed to ionizing radiation. *Journal of Radiation Research*, 50 (1):61–71, 2009. (cited on pages 35 and 67)

X. Fan, L. Shi, H. Fang, Y. Cheng, R. Perkins, and W. Tong. DNA microarrays are predictive of cancer prognosis: a re-evaluation. *Clinical Cancer Research*, 16(2):629–636, 2010. (cited on pages 47, 50, 68 and 76)

P. Fei and W. S. El-Deiry. P53 and radiation responses. *Oncogene*, 22(37):5774–5783, 2003. (cited on pages 36 and 58)

T. Fliedner, I. Friesecke, and K. Beyrer. *Medical management of radiation accidents: manual on the acute radiation syndrome*. British Institute of Radiology, 2001. (cited on page 13)

A. J. Fornace. Mammalian genes induced by radiation; activation of genes associated with growth control. *Annual Review of Genetics*, 26:507–526, 1992. (cited on page 15)

G. Gruel, C. Lucchesi, A. Pawlik, V. Frouin, O. Alibert, T. Kortulewski, A. Zarour, B. Jacquelin, X. Gidrol, and D. Tronik-Le Roux. Novel microarray-based method for estimating exposure to ionizing radiation. *Radiation Research*, 166(5):746–756, 2006. (cited on page 66)

M. Harbers and P. Carninci. Tag-based approaches for transcriptome research and genome annotation. *Nature Methods*, 2(7):495–502, 2005. (cited on page 5)

A. C. Haury, L. Jacob, and J. P. Vert. Increasing stability and interpretability of gene expression signatures. *arXiv preprint arXiv:1001.3109*, 2010. (cited on page 76)

A. C. Haury, P. Gestraud, and J. P. Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE*, 6(12):e28210, 2011. (cited on pages 47, 50 and 76)

Z. He and W. Yu. Stable feature selection for biomarker discovery. *Computational Biology and Chemistry*, 34(4):215–225, 2010. (cited on pages 44, 68, 70 and 76)

E. S. Helton and X. Chen. p53 modulation of the DNA damage response. *Journal of Cellular Biochemistry*, 100(4):883–896, 2007. (cited on page 36)

R. A. Holt and S. J. Jones. The new paradigm of flow cell sequencing. *Genome Research*, 18 (6):839–846, 2008. (cited on page 5)

d. a. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, 2009a. (cited on page 32)

D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2009b. (cited on page 32)

W. Huda and K. M. Ogden. Comparison of head and body organ doses in CT. *Physics in Medicine and Biology*, 53(2):N9–N14, 2008. (cited on page 75)

S. P. Jackson. Sensing and repairing DNA double-strand breaks. *Carcinogenesis*, 23(5): 687–696, 2002. (cited on pages 28 and 29)

A. K. Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8): 651–666, 2010. (cited on page 6)

P. Jaluria, K. Konstantopoulos, M. Betenbaugh, and J. Shiloach. A perspective on microarrays: current applications, pitfalls, and potential uses. *Microbial Cell Factories*, 6:4, 2007. (cited on page 5)

A. Jayaraman and J. Hahn. *Methods in bioengineering: systems analysis of biological networks*. Artech House Publishers, 2009. (cited on page 4)

K. Y. Jen and V. G. Cheung. Transcriptional response of lymphoblastoid cells to ionizing radiation. *Genome Research*, 13(9):2092–2100, 2003. (cited on pages 33, 40 and 42)

Y. W. Jin, Y. J. Na, Y. J. Lee, S. An, J. E. Lee, M. Jung, H. Kim, S. Y. Nam, C. S. Kim, K. H. Yang, S. U. Kim, W. K. Kim, W. Y. Park, K. Y. Yoo, C. S. Kim, and J. H. Kim. Comprehensive analysis of time- and dose-dependent patterns of gene expression in a human mesenchymal stem cell line exposed to low-dose ionizing radiation. *Oncology Reports*, 19 (1):135–144, 2008. (cited on pages 67 and 74)

G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *Proceedings of the eleventh International Conference on Machine Learning (ICML)*, volume 94, pages 121–129, 1994. (cited on page 7)

S. Kabacik, A. Mackay, N. Tamber, G. Manning, P. Finnon, F. Paillier, A. Ashworth, S. Bouffler, and C. Badie. Gene expression following ionising radiation: identification of biomarkers for dose estimation and prediction of individual response. *International Journal of Radiation Biology*, 87(2):115–129, 2011. (cited on page 67)

M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(Database issue):D109–D114, 2012. (cited on pages 7 and 32)

C. M. Kang, K. P. Park, J. E. Song, D. I. Jeoung, C. K. Cho, T. H. Kim, S. Bae, S. J. Lee, and Y. S. Lee. Possible biomarkers for ionizing radiation exposure in human peripheral blood lymphocytes. *Radiation Research*, 159(3):312–119, 2003. (cited on page 59)

T. Kawase, R. Ohki, T. Shibata, S. Tsutsumi, N. Kamimura, J. Inazawa, T. Ohta, H. Ichikawa, H. Aburatani, F. Tashiro, and Y. Taya. PH domain-only protein PHLDA3 is a p53-regulated repressor of akt. *Cell*, 136(3):535–550, 2009. (cited on page 57)

C. M. Kendziorski, Y. Zhang, H. Lan, and A. D. Attie. The efficiency of pooling mRNA in microarray experiments. *Biostatistics*, 4(3):465–477, 2003. (cited on page 68)

K. K. Khanna and S. P. Jackson. DNA double-strand breaks: signaling, repair and the cancer connection. *Nature Genetics*, 27(3):247–254, 2001. (cited on pages 28 and 29)

S. Y. Kim. Effects of sample size on robustness and prediction accuracy of a prognostic gene signature. *BMC Bioinformatics*, 10:147, 2009. (cited on pages 44 and 70)

K. A. Kirou, C. Lee, S. George, K. Louca, M. G. Peterson, and M. K. Crow. Activation of the interferon-alpha pathway identifies a subgroup of systemic lupus erythematosus patients with distinct serologic features and active disease. *Arthritis & Rheumatism*, 52(5):1491–1503, 2005. (cited on page 38)

E. Kis, T. Szatmári, M. Keszei, R. Farkas, O. Ésik, K. Lumniczky, A. Falus, and G. Sáfrány. Microarray analysis of radiation response genes in primary human fibroblasts. *International Journal of Radiation Oncology, Biology, Physics*, 66(5):1506–1514, 2006. (cited on pages 67 and 68)

K. Knops. *Genexpressionsanalysen in humanen peripheren Blutlymphozyten nach Bestrahlung - Grundlagen für biodosimetrische Applikationen*. PhD thesis, Universität Duisburg-Essen, 2013. (cited on pages 34, 42 and 75)

K. Knops, S. Boldt, O. Wolkenhauer, and R. Kriehuber. Gene expression in low- and high-dose-irradiated human peripheral blood lymphocytes: Possible applications for biodosimetry. *Radiation Research*, 178(4):304–312, 2012. (cited on page 64)

K. L. Koenig, R. E. Goans, R. J. Hatchett, F. A. Mettler, T. A. Schumacher, E. K. Noji, and D. G. Jarrett. Medical treatment of radiological casualties: current concepts. *Annals of Emergency Medicine*, 45(6):643–652, 2005. (cited on pages 12 and 13)

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence (IJCAI)*, volume 14, pages 1137–1145, 1995. (cited on page 10)

S. B. Kotsiantis. Supervised machine learning: A review of classification techniques. In *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, pages 3–24. IOS Press, 2007. (cited on page 8)

S. Lagerwerf, M. G. Vrouwe, R. M. Overmeer, M. I. Fousteri, and L. H. Mullenders. DNA damage response and transcription. *DNA Repair (Amst)*, 10(7):743–750, 2011. (cited on page 29)

P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, et al. Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1):86–112, 2006. (cited on page 8)

J. W. Lee, J. B. Lee, M. Park, and S. H. Song. An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis*, 48(4):869–885, 2005. (cited on page 8)

K. J. Lindsay, P. J. Coates, S. A. Lorimore, and E. G. Wright. The genetic basis of tissue responses to ionizing radiation. *British Journal of Radiology*, 80(Special Issue 1):S2–S6, 2007. (cited on page 36)

G. Liu and X. Chen. The ferredoxin reductase gene is regulated by the p53 family and sensitizes cells to oxidative stress-induced apoptosis. *Oncogene*, 21(47):7195–7204, 2002. (cited on pages 57 and 63)

A. Léonard, J. Rueff, G. B. Gerber, and E. D. Léonard. Usefulness and limits of biological dosimetry based on cytogenetic methods. *Radiation Protection Dosimetry*, 115(1-4):448–454, 2005. (cited on page 15)

M. Mallya, R. D. Campbell, and B. Aguado. Transcriptional analysis of a novel cluster of ly-6 family members in the human and mouse major histocompatibility complex: five genes with many splice forms. *Genomics*, 80(1):113–123, 2002. (cited on page 63)

G. Manning, S. Kabacik, P. Finnon, S. Bouffler, and C. Badie. High and low dose responses of transcriptional biomarkers in ex vivo X-irradiated human blood. *International Journal of Radiation Biology*, 89(7):512–522, 2013. (cited on page 67)

J. E. McDermott, J. Wang, H. Mitchell, B. J. Webb-Robertson, R. Hafen, J. Ramey, and K. D. Rodland. Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. *Expert opinion on medical diagnostics*, 7(1):37–51, 2013. (cited on page 3)

J. A. Meador, S. A. Ghandhi, and S. A. Amundson. p53-independent downregulation of histone gene expression in human cell lines by high- and low-LET radiation. *Radiation Research*, 175(6):689–699, 2011. (cited on pages 38 and 40)

S. K. Meadows, H. K. Dressman, G. G. Muramoto, H. Himburg, A. Salter, Z. Wei, G. S. Ginsburg, G. Ginsburg, N. J. Chao, J. R. Nevins, and J. P. Chute. Gene expression signatures of radiation response are specific, durable and accurate in mice and humans. *PLoS ONE*, 3(4):e1912, 2008. (cited on page 66)

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010. (cited on page 51)

S. Michiels, S. Koscielny, and C. Hill. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, 365(9458):488–492, 2005. (cited on page 44)

M. Mori, M. A. Benotmane, I. Tirone, E. L. Hooghe-Peters, and C. Desaintes. Transcriptional response to ionizing radiation in lymphocyte subsets. *Cellular and Molecular Life Sciences*, 62(13):1489–1501, 2005. (cited on pages 38, 42 and 67)

W. G. Nelson and M. B. Kastan. DNA strand breaks: the DNA template alterations that trigger p53-dependent DNA damage response pathways. *Molecular and Celluar Biology*, 14(3):1815–1823, 1994. (cited on page 36)

J. H. Oh, H. P. Wong, X. Wang, and J. O. Deasy. A bioinformatics filtering strategy for identifying radiation response biomarker candidates. *PloS ONE*, 7(6):e38870, 2012. (cited on page 69)

D. A. Okar and A. J. Lange. Fructose-2,6-bisphosphate and control of carbohydrate metabolism in eukaryotes. *Biofactors*, 10(1):1–14, 1999. (cited on page 63)

S. Paul and S. A. Amundson. Development of gene expression signatures for practical radiation biodosimetry. *International Journal of Radiation Oncology, Biology, Physics*, 71(4):1236–1244, 2008. (cited on pages 66, 67 and 75)

S. Paul and S. A. Amundson. Gene expression signatures of radiation exposure in peripheral white blood cells of smokers and non-smokers. *International Journal of Radiation Biology*, 87(8):791–801, 2011. (cited on page 75)

S. Paul, C. A. Barker, H. C. Turner, A. McLane, S. L. Wolden, and S. A. Amundson. Prediction of in vivo radiation dose status in radiotherapy patients using ex vivo and in vivo gene expression signatures. *Radiation Research*, 175(3):257–265, 2011. (cited on page 75)

X. Peng, C. L. Wood, E. M. Blalock, K. C. Chen, P. W. Landfield, and A. J. Stromberg. Statistical implications of pooling RNA samples for microarray experiments. *BMC Bioinformatics*, 4(1):26, 2003. (cited on page 68)

Y. Piao, M. Piao, K. Park, and K. H. Ryu. An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data. *Bioinformatics*, 28(24):3306–3315, 2012. (cited on page 51)

M. M. Pinto, N. F. Santos, and A. Amaral. Current status of biodosimetry based on standard cytogenetic methods. *Radiation and Environmental Biophysics*, 49(4):567–581, 2010. (cited on pages 13, 14 and 15)

E. Podgoršak. *Radiation oncology physics: a handbook for teachers and students*. STI/PUB. International Atomic Energy Agency, 2005. ISBN 9789201073044. (cited on page 28)

E. L. Pogosova-Agadjanyan, W. Fan, G. E. Georges, J. L. Schwartz, C. M. Kepler, H. Lee, A. L. Suchanek, M. R. Cronk, A. Brumbaugh, J. H. Engel, M. Yukawa, L. P. Zhao, S. Heimfeld, and D. L. Stirewalt. Identification of radiation-induced expression changes in nonimmortalized human T cells. *Radiation Research*, 175(2):172–184, 2011. (cited on pages 33, 42, 59 and 67)

M. Port, C. Boltze, Y. Wang, B. Röper, V. Meineke, and M. Abend. A radiation-induced gene signature distinguishes post-Chernobyl from sporadic papillary thyroid cancers. *Radiation Research*, 168(6):639–649, 2007. (cited on page 68)

J. Quackenbush. Microarray data normalization and transformation. *Nature genetics*, 32: 496–501, 2002. (cited on page 6)

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org. (cited on page 18)

S. Rana, R. Kumar, S. Sultana, and R. K. Sharma. Radiation-induced biomarkers for the detection and assessment of absorbed radiation doses. *Journal of Pharmacy & Bioallied Sciences*, 2(3):189–196, 2010. (cited on page 12)

R. B. Rao, G. Fung, and R. Rosales. On the dangers of cross-validation. an experimental evaluation. In *SIAM Data Mining*, pages 588–596, 2008. (cited on page 10)

S. Rashi-Elkeles, R. Elkon, S. Shavit, Y. Lerenthal, C. Linhart, A. Kupershtein, N. Amariglio, G. Rechavi, R. Shamir, and Y. Shiloh. Transcriptional modulation induced by ionizing radiation: p53 remains a central player. *Molecular Oncology*, 5(4):336–348, 2011. (cited on pages 29, 36, 38 and 42)

T. Rich, R. L. Allen, and A. H. Wyllie. Defying death after DNA damage. *Nature*, 407(6805): 777–783, 2000. (cited on pages 29 and 35)

J. C. Rockett and G. M. Hellmann. Confirming microarray data – is it really necessary? *Genomics*, 83(4):541–549, 2004. (cited on page 59)

H. Romm, U. Oestreicher, and U. Kulka. Cytogenetic damage analysed by the dicentric assay. *Annali dell'Istituto Superiore di Sanità*, 45(3):251–259, 2009. (cited on page 75)

Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007. (cited on pages 7 and 47)

T. Sanavia, F. Aiolli, G. Da San Martino, A. Bisognin, and B. Di Camillo. Improving biomarker list stability by integration of biological knowledge in the learning process. *BMC Bioinformatics*, 13(Suppl 4):S22, 2012. (cited on page 76)

E. Schmitt, C. Paquet, M. Beauchemin, and R. Bertrand. DNA-damage response network at the crossroads of cell-cycle checkpoints, cellular senescence and apoptosis. *Journal of Zhejiang University Science B*, 8(6):377–397, 2007. (cited on page 29)

A. J. Sigurdson, M. Ha, M. Hauptmann, P. Bhatti, R. J. Sram, O. Beskid, E. J. Tawn, C. A. Whitehouse, C. Lindholm, M. Nakano, Y. Kodama, N. Nakamura, I. Vorobtsova, U. Oestreicher, G. Stephan, L. C. Yong, M. Bauchinger, E. Schmid, H. W. Chung, F. Darroudi, L. Roy, P. Voisin, J. F. Barquinero, G. Livingston, D. Blakey, I. Hayata, W. Zhang, C. Wang, L. M. Bennett, L. G. Littlefield, A. A. Edwards, R. A. Kleinerman, and J. D. Tucker. International study of factors affecting human chromosome translocations. *Mutation Research*, 652(2):112–121, 2008. (cited on page 15)

S. L. Simon, A. Bouville, and R. Kleinerman. Current use and future needs of biodosimetry in studies of long-term health risk following radiation exposure. *Health Physics*, 98(2): 109–117, 2010. (cited on pages 11, 12 and 75)

G. K. Smyth and T. Speed. Normalization of cDNA microarray data. *Methods*, 31(4):265–273, 2003. (cited on pages 6 and 26)

C. Su, G. Gao, S. Schneider, C. Helt, C. Weiss, M. A. O'Reilly, D. Bohmann, and J. Zhao. DNA damage induces downregulation of histone gene expression through the G1 checkpoint pathway. *The EMBO journal*, 23(5):1133–1143, 2004. (cited on page 40)

W. Sudprasert, P. Navasumrit, and M. Ruchirawat. Effects of low-dose gamma radiation on dna damage, chromosomal aberration and expression of repair genes in human blood cells. *International Journal of Hygiene and Environmental Health*, 209(6):503–511, 2006. (cited on page 74)

J. M. Taylor, D. P. Ankerst, and R. R. Andridge. Validation of biomarker-based risk prediction models. *Clinical Cancer Research*, 14(19):5977–5983, 2008. (cited on page 68)

J. W. Tukey. Comparing individual means in the analysis of variance. *Biometrics*, 5(2): 99–114, 1949. (cited on page 32)

S. Varma and R. Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1):91, 2006. (cited on page 70)

R. S. Vasan. Biomarkers of cardiovascular disease molecular basis and practical considerations. *Circulation*, 113(19):2335–2362, 2006. (cited on page 3)

V. E. Velculescu, L. Zhang, W. Zhou, J. Vogelstein, M. A. Basrai, D. E. Bassett, P. Hieter, B. Vogelstein, and K. W. Kinzler. Characterization of the yeast transcriptome. *Cell*, 88 (2):243–251, 1997. (cited on page 4)

J. Vogt, N. V. Morgan, T. Marton, S. Maxwell, B. J. Harrison, D. Beeson, and E. R. Maher. Germline mutation in DOK7 associated with fetal akinesia deformation sequence. *Journal of Medical Genetics*, 46(5):338–340, 2009. (cited on page 58)

P. Voisin, L. Roy, and M. Benderitter. Why can't we find a better biological indicator of dose? *Radiation Protection Dosimetry*, 112(4):465–469, 2004. (cited on page 75)

A. Vral, M. Fenech, and H. Thierens. The micronucleus assay as a biological dosimeter of in vivo ionising radiation exposure. *Mutagenesis*, 26(1):11–17, 2011. (cited on page 14)

Q. Wang, Y. Chen, F. Xie, Y. Ge, X. Wang, and X. Zhang. A novel agonist anti-human OX40L monoclonal antibody that stimulates T cell proliferation and enhances cytokine secretion. *Hybridoma (Larchmt)*, 28(4):269–276, 2009a. (cited on page 58)

Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009b. (cited on page 4)

J. K. Waselenko, T. J. MacVittie, W. F. Blakely, N. Pesik, A. L. Wiley, W. E. Dickerson, H. Tsu, D. L. Confer, C. N. Coleman, T. Seed, P. Lowry, J. O. Armitage, and N. Dainiak. Medical management of the acute radiation syndrome: recommendations of the strategic national stockpile radiation working group. *Annals of Internal Medicine*, 140(12):1037–1051, 2004. (cited on pages 12 and 13)

A. Wojcik, E. Gregoire, I. Hayata, L. Roy, S. Sommer, G. Stephan, and P. Voisin. Cytogenetic damage in lymphocytes for the purpose of dose reconstruction: a review of three recent radiation accidents. *Cytogenetic and Genome Research*, 104(1-4):200–205, 2004. (cited on page 14)

G. Woodside. *Environmental, safety, and health engineering.* John Wiley & Sons, 1997. (cited on pages 11 and 12)

A. J. Wyrobek, C. F. Manohar, V. V. Krishnan, D. O. Nelson, M. R. Furtado, M. S. Bhattacharya, F. Marchetti, and M. A. Coleman. Low dose radiation response curves, networks and pathways in human lymphoblastoid cells exposed from 1 to 10 cGy of acute gamma radiation. *Mutation Research*, 722(2):119–130, 2011. (cited on page 67)

M. Zahurak, G. Parmigiani, W. Yu, R. B. Scharpf, D. Berman, E. Schaeffer, S. Shabbeer, and L. Cope. Pre-processing Agilent microarray data. *BMC Bioinformatics*, 8(1):142, 2007. (cited on page 6)

Supporting information for chapter 4

Table A.1: List of all radiation-responsive genes identified by the fold-change-driven gene selection for the low dose range. With the p-value-driven and consecutively applied fold-change-driven gene selection 101 genes were selected for the low dose range (0, 0.02, 0.1 Gy). The frequency of selection indicates how often a gene is among the 20 selected genes with maximal fold-change within 100 repeated 6-fold cross-validations.

Agilent ID	Gene Name	Frequency of Selections
A_24_P506680	*THC2535753*	600
A_24_P375205	*MKL2*	600
A_24_P332081	*JAKMIP3*	600
A_23_P38154	*FDXR*	600
A_24_P111096	*PFKFB3*	599
A_32_P138939	N/A	581
A_24_P341000	*FLJ35379*	544
A_32_P20997	N/A	541
A_24_P7584	*LY6G5C*	532
A_24_P392723	*CROCCL2*	481
A_24_P923510	N/A	479
A_32_P43914	N/A	430
A_32_P185741	N/A	411
A_24_P101651	N/A	367
A_24_P42308	*FLJ31887*	366
A_24_P928031	N/A	345
A_23_P102071	*FLJ14409*	341
A_32_P85230	*ISG20L1*	281
A_24_P557019	*BF476310*	263
A_24_P253454	*RGMA*	243
A_32_P7974	*TDRD10*	190
A_23_P162449	*SRGAP1*	166
A_24_P499481	N/A	161
A_24_P348594	N/A	149
A_23_P104594	*SCT*	21
A_23_P52610	*DDB2*	21
A_23_P356041	*SPAG9*	19
A_24_P310256	*LGI4*	19
A_23_P38757	*SLC14A1*	19
A_23_P129695	*VASN*	18
A_23_P4212	*HOXB13*	18
A_23_P110624	*CTNND2*	17
A_32_P179199	*BE181768*	16
A_32_P221799	*HIST1H2AM*	16
A_24_P838797	*BC042064*	15
A_24_P936470	*AF007193*	14
A_24_P678056	N/A	14
A_24_P452024	N/A	13
A_24_P361480	*TANC2*	13
A_32_P79313	*BC040596*	11
A_32_P60632	*BE175081*	11
A_23_P6481	*TNRC6B*	10
A_24_P120537	*SH3RF2*	10
A_24_P749374	*BF983943*	8
A_23_P217178	*MAGEC1*	7
A_23_P115652	*SFTPA1*	7
A_24_P10731	*MADCAM1*	7
A_23_P143673	*RASD2*	6

A_24_P147242	*A2BP1*	132
A_23_P356646	*TEKT4P2*	112
A_32_P133840	*TMCC2*	103
A_24_P490704	N/A	100
A_24_P85258	*KIAA1751*	100
A_24_P943922	*CACHD1*	95
A_23_P421526	*ODF4*	94
A_23_P54736	*GNG13*	83
A_32_P151152	N/A	79
A_23_P382775	*BBC3*	77
A_23_P113777	*ITGBL1*	76
A_24_P409182	N/A	75
A_32_P3914	N/A	74
A_32_P36767	N/A	59
A_24_P238819	N/A	57
A_23_P10640	*ENPP7*	53
A_24_P237936	*TCF23*	52
A_23_P148852	*AD7C-NTP*	45
A_23_P35055	*NPHS2*	38
A_24_P901778	*BM926140*	36
A_23_P31858	*ST18*	34
A_24_P102650	*MUC5B*	34
A_32_P116538	N/A	32
A_23_P58642	*PITX1*	28
A_32_P208599	*AB074268*	26
A_32_P149404	*LOC728371*	24
A_24˙P281514	*LOC730589*	6
A_32_P99804	N/A	6
A_24_P10751	*HNF4A*	5
A_23_P72411	*CYP4X1*	5
A_23_P301971	*MASP2*	5
A_23_P59691	*PAX4*	5
A_32_P459423	*FUNDC2*	4
A_24_P373877	*SYT5*	4
A_23_P5995	*ARFGEF2*	3
A_23_P501713	*IL1F10*	3
A_23_P108534	N/A	3
A_24_P289665	*LOC389332*	3
A_24_P99679	*POLR2A*	3
A_23_P169007	*NPM2*	3
A_24_P195272	*NP450512*	3
A_32_P60707	N/A	2
A_23_P12128	*TSHB*	2
A_32_P157775	*THC2564488*	2
A_24_P281403	*OR4C46*	2
A_23_P70733	*TAAR2*	2
A_32_P405703	*KIAA116*1	2
A_24_P809964	*AA465699*	2
A_24_P388593	*MAP6D1*	2
A_23_P217704	*GYG2*	1
A_23_P131289	*CHPF*	1
A_24_P552677	N/A	1
A_24_P187218	*PCDH9*	1

Table A.2: List of radiation-responsive genes 24 h after low dose exposure (0 Gy, 0.02 Gy, 0.1 Gy).

		0 Gy			**0.02 Gy**			**0.1 Gy**		
Agilent ID	**Gene Name**	**Rep1**	**Rep2**	**Rep3**	**Rep1**	**Rep2**	**Rep3**	**Rep1**	**Rep2**	**Rep3**
A_23_P54736	*GNG13*	9.12	8.78	8.67	9.89	9.68	10.36	10.44	10.76	10.05
A_23_P162449	*SRGAP1*	5.99	6.36	5.99	6.96	6.94	7.47	7.19	7.90	7.56
A_32_P149404	*LOC728371*	4.77	4.44	4.65	5.89	5.30	5.46	6.19	6.14	5.61
A_32_P3914	N/A	8.73	8.32	8.14	9.27	9.01	9.70	9.77	10.00	9.46
A_32_P30421	*AA975908*	5.40	5.09	5.13	6.13	6.02	6.52	6.58	6.82	6.24
A_23_P38154	*FDXR*	11.56	11.57	11.46	12.01	12.01	12.50	13.18	12.46	12.93
A_32_P231143	N/A	4.90	4.39	4.33	5.82	5.58	5.18	6.00	5.88	5.70
A_23_P356646	*LOC389833*	8.33	8.02	8.08	8.99	8.55	9.27	9.56	9.62	9.15
A_32_P43914	*THC2712687*	9.34	9.21	9.01	9.95	10.01	10.22	10.61	10.61	10.14
A_32_P189324	N/A	5.85	5.45	5.31	6.52	5.92	6.60	6.51	6.96	6.80
A_23_P169007	*NPM2*	6.86	6.55	6.94	7.63	7.50	7.90	8.09	8.15	7.69
A_24_P237936	*TCF23*	7.40	7.46	7.37	8.20	7.85	8.49	8.61	8.71	8.41
A_24_P901778	*BM926140*	5.94	5.43	5.53	6.40	6.00	6.58	7.00	6.76	6.55
A_32_P138939	N/A	6.64	6.51	6.32	7.03	6.75	7.51	7.79	7.44	7.53
A_23_P430120	*EPAS1*	5.74	6.06	5.67	6.28	6.51	6.51	7.13	6.91	6.68
A_24_P110887	*KIAA2018*	6.24	6.10	6.42	5.51	5.37	5.31	5.10	5.15	5.27
A_32_P85230	*BE646426*	9.78	9.60	9.58	9.97	9.94	9.92	10.78	10.72	10.63
A_32_P208599	*AB074268*	7.13	7.22	6.75	7.61	7.68	7.73	8.14	8.35	7.70
A_23_P131289	*CHPF*	6.17	6.18	6.11	6.71	6.85	6.73	7.38	7.31	6.86
A_23_P17914	*PNPLA3*	5.26	5.62	5.51	4.96	4.76	4.65	4.28	4.44	4.60
A_23_P23885	*GJB4*	5.01	5.27	4.85	5.67	5.38	5.82	6.09	6.22	5.87
A_23_P148852	*AD7C-NTP*	7.16	6.90	6.90	7.58	7.40	7.83	7.99	8.29	7.71
A_23_P104224	*ACF*	5.15	4.63	4.73	5.71	5.15	5.44	5.90	6.19	5.45
A_23_P206077	*ISG20L1*	9.10	9.43	9.30	9.59	9.80	9.48	10.39	10.06	10.39

Table A.3: List of significant radiation-responsive genes 48 h after low dose exposure (0 Gy, 0.02 Gy, 0.1 Gy).

		0 Gy			0.02 Gy			0.1 Gy		
Agilent ID	**Gene Name**	**Rep1**	**Rep2**	**Rep3**	**Rep1**	**Rep2**	**Rep3**	**Rep1**	**Rep2**	**Rep3**
A_23_P200425	*S100PBP*	10.31	10.31	10.28	10.70	10.81	10.84	10.47	10.41	10.49
A_24_P33048	*C14orf28*	9.58	9.58	9.58	9.90	9.90	9.88	9.84	9.71	9.81
A_24_P85590	*AF116684*	4.76	4.76	4.52	5.42	5.57	5.62	5.41	5.41	5.23
A_23_P152038	*TUBGCP5*	8.87	8.87	8.85	8.26	8.15	8.21	8.55	8.49	8.69
A_32_P208599	*AB074268*	7.01	7.01	6.78	8.22	8.34	8.54	8.12	8.10	8.07
A_32_P170454	*LOC283454*	8.09	8.09	8.10	8.58	8.78	8.76	8.10	8.20	8.10
A_24_P42001	*LOC728044*	6.82	6.82	6.76	7.91	7.86	7.89	7.58	7.50	7.66
A_32_P138939	N/A	6.49	6.49	6.32	7.81	7.71	7.38	7.46	7.73	7.60
A_23_P22473	*FTSJ1*	9.60	9.60	9.54	9.19	9.20	9.15	9.53	9.47	9.55
A_23_P393697	*MGC27345*	10.35	10.35	10.27	10.61	10.61	10.69	10.43	10.47	10.45
A_24_P325533	N/A	5.00	5.00	4.88	6.15	6.05	5.97	5.63	5.85	5.92
A_23_P417383	*SASP*	10.30	10.30	10.39	11.07	11.40	11.39	10.63	10.70	10.77
A_32_P113472	*CA314451*	11.06	11.06	10.90	11.97	11.90	12.27	11.35	11.36	11.41
A_23_P77965	*ABC1*	10.09	10.09	9.98	9.65	9.68	9.63	9.88	9.87	9.86
A_23_P114689	*DDEFL1*	5.47	5.47	5.61	6.58	6.46	6.56	6.24	5.90	6.14
A_23_P147121	*ZNF136*	9.76	9.76	9.82	9.36	9.25	9.35	9.63	9.53	9.53
A_32_P96692	*THC2532504*	9.52	9.52	9.60	10.16	10.17	10.31	9.97	10.02	10.09
A_24_P238819	N/A	7.29	7.29	7.04	8.22	8.32	8.19	7.94	8.11	8.04
A_23_P253586	*DOPEY2*	9.43	9.43	9.41	8.73	8.90	8.89	9.19	9.10	9.09
A_23_P211785	*ZNF35*	8.12	8.12	8.12	7.48	7.64	7.40	7.88	7.81	7.84
A_32_P116538	*THC2686110*	6.72	6.72	6.52	8.04	7.56	7.91	7.67	7.63	7.75
A_23_P24716	*TMEM132A*	6.40	6.40	6.40	6.77	6.81	6.74	6.55	6.49	6.57
A_32_P36767	N/A	5.96	5.96	6.10	7.20	6.81	6.90	6.81	6.83	6.90
A_23_P256384	*PSIP1*	9.20	9.20	9.21	9.80	9.94	10.02	9.41	9.37	9.58

		0 Gy			0.02 Gy			0.1 Gy		
Agilent ID	**Gene Name**	**Rep1**	**Rep2**	**Rep3**	**Rep1**	**Rep2**	**Rep3**	**Rep1**	**Rep2**	**Rep3**
A_24_P332911	N/A	7.09	7.09	7.08	8.05	7.95	8.10	7.67	7.84	7.69
A_24_P474188	N/A	9.21	9.21	9.17	8.72	8.59	8.68	9.08	8.93	8.94
A_32_P87631	*BC017507*	9.96	9.96	9.97	10.30	10.45	10.44	10.04	9.93	9.93
A_32_P208621	*THC2687042*	5.20	5.20	5.09	5.96	5.97	5.97	6.14	5.94	5.85
A_23_P374104	*ANGPTL2*	5.31	5.31	5.23	6.68	6.96	7.20	6.17	6.57	6.46
A_23_P259333	*C6orf203*	11.55	11.55	11.49	11.93	11.86	11.99	11.67	11.74	11.72
A_24_P134863	*ITIH5L*	4.90	4.90	4.87	5.97	5.89	5.99	5.79	5.69	5.52
A_32_P134657	*KIAA0240*	13.13	13.13	13.08	13.84	14.24	14.15	13.44	13.51	13.49
A_24_P300841	*CKAP5*	9.21	9.21	9.22	8.52	8.59	8.47	8.99	8.75	8.90
A_23_P104065	*ICMT*	9.81	9.81	9.83	9.23	9.43	9.27	9.60	9.63	9.61
A_23_P88106	*MCF2L*	7.67	7.67	7.69	7.26	7.20	7.11	7.60	7.49	7.49
A_32_P20997	*BU561469*	10.21	10.21	10.10	11.49	11.72	11.40	11.33	11.61	11.43
A_23_P416581	*GNAZ*	8.26	8.26	8.13	9.18	9.20	9.19	8.64	8.55	8.67
A_24_P310864	*CCDC24*	8.46	8.46	8.39	8.79	8.89	8.86	8.49	8.54	8.53
A_23_P410982	*LOC392473*	9.30	9.30	9.29	8.82	8.63	8.60	9.15	9.21	9.13
A_23_P157283	*C7orf23*	12.51	12.51	12.58	12.90	12.91	12.98	12.74	12.68	12.79
A_32_P184746	*THC2715480*	10.45	10.45	10.41	11.03	11.23	11.07	10.69	10.57	10.49
A_24_P464798	N/A	13.89	13.89	13.86	14.21	14.18	14.25	14.05	13.96	14.04
A_23_P88435	*CHES1*	11.19	11.19	11.09	10.77	10.76	10.76	10.94	10.82	10.87
A_32_P148058	*RUNX2*	5.78	5.78	5.77	6.91	6.86	6.66	6.54	6.60	6.49
A_24_P177844	*ENST00000283694*	11.06	11.06	11.01	11.65	11.76	11.73	11.34	11.36	11.41
A_24_P241996	*AM181370*	6.60	6.60	6.54	7.45	7.47	7.21	7.14	7.15	7.18
A_23_P38876	*LIPE*	10.25	10.25	10.24	10.91	10.81	10.77	10.55	10.58	10.73
A_24_P298545	*ZNF257*	11.26	11.26	11.24	11.55	11.51	11.51	11.41	11.33	11.32
A_32_P159726	*KIAA1244*	8.92	8.92	8.79	9.69	9.92	9.65	9.27	9.41	9.46
A_24_P260101	*MME*	6.44	6.44	6.44	5.54	5.65	5.66	5.89	5.84	5.93

		0 Gy			0.02 Gy			0.1 Gy		
Agilent ID	**Gene Name**	**Rep1**	**Rep2**	**Rep3**	**Rep1**	**Rep2**	**Rep3**	**Rep1**	**Rep2**	**Rep3**
A_24_P383199	*ENST00000260303*	6.80	6.80	6.53	7.92	8.03	7.63	7.52	7.49	7.56
A_24_P315885	N/A	8.01	8.01	7.79	9.19	9.11	9.15	8.73	8.91	8.73
A_23_P59481	*UBE3C*	11.02	11.02	10.98	10.75	10.71	10.71	10.84	10.90	10.94
A_24_P120537	*SH3RF2*	4.96	4.96	4.60	6.25	6.14	6.00	5.80	5.73	6.01
A_24_P296280	*FAM82C*	9.92	9.92	9.98	9.66	9.67	9.60	9.83	9.78	9.76
A_23_P216894	*MAPKAP1*	9.54	9.54	9.47	8.86	8.85	8.84	9.44	9.23	9.30
A_23_P53081	*OSBPL5*	12.62	12.62	12.65	13.05	13.11	13.20	12.65	12.79	12.76
A_23_P308552	*DLX3*	7.41	7.41	7.17	8.40	8.63	8.58	7.98	8.14	8.33
A_32_P82111	*LRFN2*	7.16	7.16	7.19	7.57	7.61	7.70	7.32	7.36	7.37
A_23_P6119	*SEC23B*	11.61	11.61	11.64	11.25	11.25	11.30	11.50	11.59	11.48
A_23_P6085	*SLC23A2*	8.30	8.30	8.24	7.88	7.93	7.95	8.33	8.45	8.32
A_23_P60387	*NOTCH1*	10.40	10.40	10.41	9.74	9.95	9.86	10.15	10.09	10.14
A_23_P340728	*PSEN1*	10.77	10.77	10.76	10.21	9.94	9.95	10.59	10.46	10.56
A_32_P28828	*THC2674068*	10.83	10.83	10.86	11.21	11.39	11.35	10.98	10.93	10.92
A_24_P348594	*UPK3A*	7.79	7.79	7.60	8.91	9.05	8.86	8.57	8.77	8.81
A_23_P128993	*GZMH*	12.48	12.48	12.34	13.31	13.13	13.20	12.70	12.68	12.73
A_24_P289665	*LOC389332*	5.52	5.52	5.01	7.26	6.87	7.10	6.79	6.61	6.89
A_32_P142459	*BI520341*	7.33	7.33	7.28	8.20	7.94	8.13	7.86	7.77	7.88
A_24_P201728	*C12orf32*	8.93	8.93	8.85	8.57	8.48	8.49	8.79	8.78	8.86
A_24_P367432	*IGHV1-69*	8.73	8.73	8.72	9.16	9.17	9.10	8.95	8.87	9.03
A_24_P7494	N/A	7.64	7.64	7.48	9.04	9.44	8.89	8.71	8.65	8.73
A_32_P85379	*BI869933*	5.06	5.06	5.11	5.82	5.91	5.99	5.01	5.33	5.29
A_24_P892612	*AL833309*	7.46	7.46	7.47	7.19	7.19	7.23	7.37	7.35	7.34
A_23_P368681	*GIMAP2*	13.54	13.54	13.65	13.90	13.95	13.94	13.72	13.72	13.78
A_24_P264790	*LTBP3*	7.18	7.18	7.19	7.80	7.70	7.63	7.59	7.57	7.48
A_32_P179526	*ZBTB20*	11.84	11.84	11.78	12.71	12.92	13.03	11.90	12.23	12.20

		0 Gy			0.02 Gy			0.1 Gy		
Agilent ID	**Gene Name**	**Rep1**	**Rep2**	**Rep3**	**Rep1**	**Rep2**	**Rep3**	**Rep1**	**Rep2**	**Rep3**
A_32_P71212	N/A	8.55	8.55	8.78	9.64	9.82	9.66	9.10	8.95	9.16
A_32_P31666	*SUPT3H*	8.23	8.23	8.31	8.79	8.85	8.93	8.48	8.34	8.44
A_32_P157671	N/A	8.87	8.87	8.78	8.46	8.39	8.45	8.81	8.75	8.83
A_23_P359376	*HCG4P6*	5.79	5.79	5.60	5.03	5.19	5.11	5.66	5.62	5.68
A_24_P936470	*AF007193*	4.58	4.58	4.31	5.77	5.72	5.92	5.32	5.43	5.67
A_24_P389714	*ZRANB1*	9.32	9.32	9.35	9.77	9.91	9.92	9.57	9.47	9.56
A_23_P129157	*NEIL1*	10.01	10.01	9.91	10.51	10.49	10.63	10.14	10.03	10.13
A_23_P309599	*C20orf119*	8.02	8.02	8.02	7.95	7.96	7.97	7.92	7.92	7.90
A_23_P169007	*NPM2*	6.57	6.57	6.44	8.37	8.16	8.34	7.79	7.78	7.93
A_24_P131066	*IGSF2*	8.47	8.47	8.42	7.60	7.25	7.39	8.06	8.15	8.22
A_24_P475115	N/A	12.71	12.71	12.63	12.08	11.83	11.95	12.46	12.58	12.52
A_23_P30464	*PRR7*	9.39	9.39	9.35	9.76	9.67	9.73	9.53	9.50	9.59
A_24_P196318	*SLC2A9*	7.52	7.52	7.48	7.07	6.92	6.87	7.45	7.43	7.31
A_32_P45780	*THC2622569*	5.61	5.61	5.72	6.60	6.54	6.83	6.13	6.11	6.11
A_32_P87145	*BC004503*	7.20	7.20	7.22	7.73	7.82	7.63	7.50	7.47	7.55
A_24_P175156	*CLEC5A*	8.15	8.15	8.16	7.66	7.58	7.55	7.94	8.11	8.11
A_24_P281514	*LOC730589*	7.55	7.55	7.36	8.79	8.72	8.72	8.41	8.48	8.48
A_23_P215642	*TNS3*	8.47	8.47	8.36	7.57	7.59	7.38	8.12	8.11	7.89
A_24_P358321	*ENST00000377233*	8.68	8.68	8.69	9.55	9.57	9.58	9.28	9.40	9.32
A_23_P104224	*ACF*	4.55	4.55	4.74	6.25	6.06	5.87	5.73	5.85	5.60
A_32_P134634	*THC2530077*	6.61	6.61	6.68	6.83	6.95	6.73	7.27	7.23	7.35
A_23_P398854	*DOK7*	7.61	7.61	7.44	8.17	8.36	8.13	7.65	7.56	7.54
A_24_P366670	*AFF3*	7.38	7.38	7.35	8.18	8.17	7.98	7.96	7.95	7.87
A_24_P195272	*NP450512*	7.53	7.53	7.23	8.64	8.64	8.78	8.47	8.42	8.46
A_24_P264507	*LOC729792*	5.14	5.14	5.00	5.99	6.20	5.92	5.81	5.83	5.74
A_24_P896032	*THC2542952*	6.25	6.25	6.23	6.71	6.73	6.82	6.45	6.55	6.56

		0 Gy			0.02 Gy			0.1 Gy		
Agilent ID	**Gene Name**	**Rep1**	**Rep2**	**Rep3**	**Rep1**	**Rep2**	**Rep3**	**Rep1**	**Rep2**	**Rep3**
A_23_P122375	*ZFAND3*	9.91	9.91	9.83	9.56	9.41	9.48	9.79	9.81	9.76
A_24_P678056	N/A	6.74	6.74	6.80	7.70	7.51	7.36	7.41	7.39	7.45
A_24_P195998	*ANKRD17*	10.02	10.02	10.11	9.66	9.60	9.65	9.80	9.80	9.72
A_24_P18137	*NEFL*	7.07	7.07	7.05	8.00	7.90	7.93	7.68	7.40	7.48
A_24_P254949	*PGM5*	7.19	7.19	7.17	8.29	8.00	7.98	7.89	7.91	7.86
A_32_P41267	*THC2659782*	7.68	7.68	7.72	8.03	8.07	7.98	7.85	7.87	7.88
A_32_P233457	*BQ050540*	6.59	6.59	6.56	6.20	6.22	6.15	6.34	6.25	6.39
A_23_P215109	*CPA5*	6.74	6.74	6.72	7.14	7.27	7.31	7.06	7.04	7.15
A_24_P356130	*MAP2K5*	9.01	9.01	9.02	8.48	8.43	8.48	8.74	8.85	8.83
A_23_P66766	*RPAIN*	12.24	12.24	12.29	12.62	12.75	12.66	12.44	12.50	12.56
A_32_P162939	*CB047924*	5.69	5.69	5.81	6.49	6.35	6.33	6.18	6.12	6.29
A_23_P5325	*ERCC3*	11.29	11.29	11.25	10.96	10.95	10.94	11.09	11.19	11.20
A_32_P77626	*CA866957*	6.27	6.27	6.17	6.75	6.85	6.75	6.45	6.62	6.52
A_23_P99360	*TRIM13*	11.40	11.40	11.46	12.14	12.20	12.41	11.75	11.70	11.86
A_32_P100830	*THC2686967*	8.49	8.49	8.56	8.87	8.98	9.00	8.47	8.42	8.45
A_23_P85936	*AL137472*	7.20	7.20	7.07	8.02	7.92	8.24	7.68	7.76	7.77
A_23_P113777	*ITGBL1*	5.36	5.36	5.31	6.82	6.51	6.40	6.51	6.78	6.55
A_32_P334325	*RIMBP2*	6.34	6.34	6.28	7.33	7.16	6.99	6.88	6.79	6.81
A_23_P147238	*WSB2*	10.26	10.26	10.23	9.69	9.50	9.70	10.01	9.95	10.05
A_24_P191312	*SLC1A4*	9.12	9.12	9.03	8.26	8.18	8.33	8.64	8.69	8.63
A_23_P153390	*CLEC4G*	7.53	7.53	7.42	7.86	7.95	7.84	7.78	7.78	7.81
A_32_P170621	*ENST00000360758*	6.95	6.95	6.90	7.91	7.82	8.14	7.80	7.63	7.76
A_23_P351069	*SOCS3*	6.29	6.29	6.26	5.55	5.57	5.33	5.88	5.97	5.82
A_24_P304449	*KIAA0152*	9.71	9.71	9.68	9.18	9.26	9.14	9.62	9.50	9.59
A_24_P901778	*BM926140*	5.79	5.79	5.53	6.97	6.67	6.93	6.65	6.59	6.68
A_23_P10614	*PDK1*	7.59	7.59	7.64	7.16	7.22	7.26	7.43	7.49	7.53

Agilent ID	Gene Name	0 Gy			0.02 Gy			0.1 Gy		
		Rep1	**Rep2**	**Rep3**	**Rep1**	**Rep2**	**Rep3**	**Rep1**	**Rep2**	**Rep3**
A_24_P756494	*AK057923*	4.39	4.39	4.27	5.38	5.28	5.33	5.16	4.96	5.23
A_23_P15786	*KRT25*	4.34	4.34	3.70	6.52	6.00	6.50	6.15	6.09	5.94
A_24_P307674	*LRRC41*	6.60	6.60	6.57	7.09	7.03	7.01	7.22	7.14	7.21
A_24_P483145	*LOC286272*	7.07	7.07	7.19	6.72	6.71	6.72	6.77	6.67	6.67
A_23_P142174	*FOXA3*	6.37	6.37	6.05	7.43	7.31	7.36	7.23	6.95	7.18
A_24_P592421	*NHEJ1*	7.87	7.87	7.81	8.26	8.43	8.37	7.96	7.86	7.85
A_23_P170908	*NUBPL*	7.66	7.66	7.77	7.31	7.26	7.22	7.53	7.52	7.49
A_24_P587993	N/A	7.67	7.67	7.40	8.76	8.66	8.67	8.31	8.35	8.59
A_32_P147865	*THC2647276*	6.31	6.31	6.15	7.64	7.70	7.64	7.38	7.33	7.32
A_23_P311640	*HRBL*	9.10	9.10	9.17	9.56	9.47	9.47	9.32	9.37	9.37
A_24_P466590	*BM692484*	8.71	8.71	8.72	8.51	8.53	8.46	8.55	8.56	8.56
A_32_P99804	N/A	7.82	7.82	7.67	8.57	8.71	8.74	8.25	8.42	8.49
A_24_P135769	*LOH11CR2A*	7.10	7.10	7.03	6.27	6.15	6.26	6.84	6.61	6.62
A_32_P84526	*BX449754*	6.57	6.57	6.67	7.30	7.19	7.12	6.85	7.00	6.95
A_24_P673968	*ENST00000371276*	11.29	11.29	11.34	12.15	12.36	12.17	11.62	11.81	11.85
A_32_P220183	*THC2651178*	8.16	8.16	8.12	8.45	8.38	8.39	8.18	8.23	8.24

Theses

1. Transcriptional changes in *ex vivo* γ-irradiated human peripheral blood lymphocytes can be analyzed by gene expression profiling. (Chapter 2)

2. The statistical and functional analysis of DNA-microarray data enables the investigation of the cellular DNA damage response to ionizing radiation. (Chapter 3)

3. Both the radiation dose and the time after irradiation have a substantial impact on the number of radiation-induced genes as well as on the cell signaling pathways and biological processes activated after exposure to ionizing radiation. (Chapter 3)

4. The DNA-microarray analysis led to the hypothesis that acute low dose exposure causes DNA-lesions which are sufficient to induce apoptosis. (Chapter 3)

5. A combined statistical and bioinformatics approach allows the identification of gene expression signatures from high-throughput data functioning as candidate radiation biodosimeters. (Chapter 4)

6. Based on a potential biomarker signature, comprising seven radiation-induced genes, medium to high radiation doses can be accurately predicted within a time frame essential for medical decisions in a radiologic emergency. (Chapter 4)

7. The ability to estimate low radiation doses with an expression signature of nine genes could be used to assess the long-term health risks associated with low dose exposure. (Chapter 4)

8. A translation of the presented *in vitro* results to *in vivo* remains a challenge, but is necessary for the development of a refined and customized biodosimetry platform allowing a fast and minimally-invasive retrospective estimation of radiation doses of exposed individuals in the future. (Chapter 5)

List of publications

Peer-reviewed Journal publications and contributions

The written thesis is mainly based on two publications, which are listed in the following. I participated in conceiving and designing the research, established and implemented the computational framework for biomarker discovery and radiation dose prediction and carried out the statistical and bioinformatics-driven data analysis. All experiments have been performed by Dr. Katja Knops. Together, we discussed the (experimental and computational) results and wrote the manuscripts. Prof. Olaf Wolkenhauer and Dr. Ralf Kriehuber supervised the work and contributed to the writing of the final manuscripts.

- **Boldt S***, Knops K*, Kriehuber R, Wolkenhauer O (2012) A frequency-based gene selection method to identify robust biomarkers for radiation dose prediction. *International Journal of Radiation Biology* 3:267-276. doi:10.3109/09553002.2012.638358.
  *These authors contributed equally to this work.

- Knops K*, **Boldt S***, Wolkenhauer O, Kriehuber R (2012) Gene expression in low and high dose-irradiated human peripheral blood lymphocytes: Possible applications for biodosimetry. *Radiation Research* 178:304-312. doi:10.1667/RR2913.1.
  *These authors contributed equally to this work.

In addition to the work on gene expression-based biodosimetry, I participated in other projects from 2007 to 2010, which led to the following journal publication and book chapter.

- Schult C, Dahlhaus M, Ruck S, Sawitzky M, Amoroso F, Lange S, Etro D, Glass Ä, Füllen G, **Boldt S**, Wolkenhauer O, Neri LM, Freund M, Junghanß C (2010) The multikinase inhibitor Sorafenib displays significant antiproliferative effects and induces apoptosis via caspase 3, 7 and PARP in B- and T-lymphoblastic cells. *BMC Cancer* 10:560. doi:10.1186/1471-2407-10-560.

  Änne Glass and I performed the computational data analysis of pilot experiments conducted to conceive and design the presented research. I participated in discussing the results and approved the final manuscript.

- Assmus HE, **Boldt S**, Wolkenhauer O (2009) Reverse Engineering of Biological Networks in: Methods in Bioengineering: Systems Analysis of Biological Networks; Jayaraman A, Hahn J (eds.); Artech House 2009. ISBN: 978-1-59693-406-1.

  I had a major impact on conceiving and designing the presented review dealing with the inference of biological networks from experimental data. I performed an extensive literature search and wrote parts of the chapter. Dr. Heike Assmus coordinated the effort, wrote parts of the manuscript and approved together with Prof. Olaf Wolkenhauer the final manuscript.

Poster at international conferences

- **Strunz S**, Wolkenhauer O, Repsilber D: An ensemble-based approach to infer gene regulatory networks from expression profiles.*European Conference of Computational Biology*, Berlin (Germany), July 2013.

- **Boldt S**, Knops K, Kriehuber R, Wolkenhauer O: Identification of gene-based biosignatures for radiation biodosimetry. *58. Biometrisches Kolloquium*, Berlin (Germany), March 2012.

- Knops K, **Boldt S**, Wolkenhauer O, Kriehuber R: Identification of radiation-specific gene expression changes in human PBL after *ex vivo* irradiation suitable for biodosimetric applications. *14th International Congress of Radiation Research*, Warsaw (Poland), August 2011.

- Knops K, **Boldt S**, Wolkenhauer O, Kriehuber R: Radiation responsive genes in human lymphocytes as a tool for radiation biodosimetry.*GBS Jahrestagung*, Hamburg (Germany), September 2010.

- Knops K, **Boldt S**, Wolkenhauer O, Kriehuber R: Gene expression pattern analysis in human PBLs as a tool for radiation biodosimetry. *EPRBioDose*, Mandelieu-La-Napoule (France), October 2010.

- **Boldt S**, Knops K, Kriehuber R, Wolkenhauer O: A *p*-value- and fold-change-driven gene selection method for radiation dose prediction. *EPRBioDose*, Mandelieu-La-Napoule (France), October 2010.

- Knops K, **Boldt S**, Wolkenhauer O, Kriehuber R: Gene Expression pattern analysis as a tool for radiation biodosimetry. *37th Annual Meeting of the European Radiation Research Society*, Prague (Czech Republic), August 2009.

- **Boldt S**, Knops K, Kriehuber R, Wolkenhauer O: A data analysis workflow for the identification of expression signatures functioning as molecular biodosimeters. *German Conference of Bioinformatics*, Halle (Germany), September 2009.

- **Boldt S**, Glass Ä, Schult C, Junghanß C, Wolkenhauer O: Detecting differential gene expression in ALL cell lines treated with the multikinase inhibitor Sorafenib. *2nd summer school on Systems Biology for Medical Application*, Costa Adeje (Tenerife), October 2008.

Presentations

- Identification of biomarker signatures for radiation biodosimetry. *Projektaustauschgespräch zur BMBF geförderten Nuklearen Sicherheitsforschung*, Karlsruhe (Germany), March 2013
- Logical models for analyzing biochemical interaction networks. *Workshop on Mathematical Modelling*, Rostock (Germany), August 2012 (invited talk)
- Using logical models in Systems Medicine. A case study. *Workshop on Mathematical Modelling*, Rostock (Germany), August 2012 (invited talk)
- The role of Systems Biology and Bioinformatics within radiation research. *Project Status Meeting*, Essen (Germany), July 2011

Selbständigkeitserklärung

Ich erkläre hiermit, die vorliegende Arbeit selbstständig und ohne unerlaubte Hilfe verfasst zu haben. Ich versichere, dass ich ausschliesslich die angegebenen Quellen und Hilfsmittel verwendet habe und ich die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Rostock, 15. November 2013

Sonja Strunz