SILESIAN UNIVERSITY OF TECHNOLOGY

Faculty of Automatic Control, Electronics and Computer Science

Institute of Automatic Control

Integrative data analysis methods in multi-omics molecular biology studies for disease of affluence biomarker research

Doctoral Dissertation

by

Anna Papież

Supervisor

prof. dr hab. inż. Joanna Polańska

2019

Gliwice, POLAND

To my Family

Abstract

The need for transforming large amounts of data in the life sciences drives the development of statistical and data mining algorithms for merging and validation of biomedical experiments. Although this issue has been previously commonly acknowledged in the scientific community, the constantly increasing amounts of data require continuous efforts towards the optimization of data analysis pipelines. Therefore, the aim of this thesis is to investigate diverse approaches for high-throughput molecular biology integrative data analysis to enable the discovery of disease of affluence biomarkers. The work consists of a detailed overview of existing advancements in high-throughput molecular biology techniques data integration, followed by the demonstration of novel algorithms for combined analysis of data derived from multi-platform and multi-domain experiments.

Initially, an original batch effect identification algorithm based on dynamic programming is presented, as correcting for these effects constitutes a part of the intra-experiment data integration pipeline. Its performance on identifying batch structure is proven to be highly efficient, and moreover, batch effect preprocessing entails potential new knowledge discovery in studied diseases and conditions.

Subsequently, two microarray data sets obtained using different platforms for biomarker research in breast cancer patients are analyzed to highlight the potential of measurement transformation to achieve computational and biological consistency. The statistical and data mining integrative approaches with functional validation and profile modeling provides a comprehensive solution for elucidating dose response mechanisms and potential biomarker signatures. Moreover, custom statistical integrative methods applied to a transcriptomics and proteomics data set on ischemic heart disease plutonium mine workers enabled discrimination of dose dependent protein expression changes from the age dependent changes and validation of pathways identified previously in the proteomic data. Another approach to data integration, which enabled the identification of factors playing a key role in differentiation of irradiated samples, was conducted on multi-tissue exosome proteomics data.

Acknowledgements

I wish to thank Professor Joanna Polańska for her invaluable guidance and outstanding introduction to the academic world.

I would also like to thank Christophe Badie from Public Health England and Soile Tapio from Helmholtz Centre Munich for the opportunity to share a cooperation during projects at their institutions, biological knowledge support and their time devoted to helpful discussions.

I thank my colleagues Joanna Żyła, Franciszek Binczyk, Michał Marczyk, and Wojciech Łabaj for their helping hand and friendship, and the entire Data Mining Laboratory for creating a very special environment for scientific brainstorming enhanced with a smile.

Thank you Mom and Dad for teaching me that only a good person can be truly wise.

Thank you Maciek, my Sweetheart, for always sailing right behind.

Funding

This work was funded partially by:

- "DoktoRIS - Scholarship Program for Innovative Silesia",
  co-financed by the European Union under the European Social Fund
- POIG.02.03.01-24-099/13 grant funding and infrastructure:
  "GeCONiI - Upper Silesian Center for Computational Science and Engineering"
- National Science Centre Harmonia grant
  number DEC-2013/08/M/ST6/00924: "BioRadInt"
- National Science Centre OPUS grant
  number UMO-2015/19/B/ST6/01736: "BiTIMS"
- National Center for Research and Development PBS grant
  number PBS3/A7/29/2015/ID-247184 "MOLTEST BIS"

Contents

List of Figures

List of Tables

List of Abbreviations

BatchI	**Batch** effect **I**dentification using dynamic programming
BMI	**B**ody **M**ass **I**ndex
ComBat	**Com**bining **Bat**ches of expression data
CVD	**C**ardio**V**ascular **D**isease
DEG	**D**ifferentially **E**xpressed **G**enes
DP	**D**ynamic **P**rogramming
FDR	**F**alse **D**iscovery **R**ate
FSHD	**F**acio**S**capulo**H**umeral muscular **D**ystrophy
GO	**G**ene **O**ntology
gPCA	**g**uided **P**rincipal **C**omponent **A**nalysis
HTS	**H**igh **T**hroughput **S**equencing
IC	**I**nformation **C**ontent
IHD	**I**schemic **H**eart **D**isease
KEGG	**K**yoto **E**ncyclopedia of **G**enes and **G**enomes
LC-MS/MS	**L**iquid **C**hromatography tandem **M**ass **S**pectrometry
MCFS	**M**onte **C**arlo **F**eature **S**election
MILE	**Microarray** **I**nnovations in **LE**ukemia
MRV	**M**ultiple **R**andom **V**alidation
MS	**M**ass **S**pectrometry
NPV	**N**egative **P**redictive **V**alue
PCA	**P**rincipal **C**omponent **A**nalysis
PPV	**P**ositive **P**redictive **V**alue
QI	**Q**uality **I**ndex
qPCR	**q**uantitative **P**olymerase **C**hain **R**eaction
RMA	**R**obust **M**ultichip **A**verage

RNA-seq	**RNA-seq**uencing
RR	**R**adio**R**esistant
RS	**R**adio**S**ensitive
RUV	**R**emove **U**nwanted **V**ariation
SNP	**S**ingle **N**ucleotide **P**olymorphism
SVA	**S**urrogate **V**ariable **A**nalysis
SVM	**S**upport **V**ector **M**achine
TIC	**T**otal **I**on **C**urrent
XPN	**Cross** **P**latform **N**ormalization

Chapter 1

Introduction

1.1 Motivation

Combining information from high-throughput cellular biology data sets has become an essential task for scientific researchers. The never ceasing growth of data available through various repositories implies the urge of raising the processing algorithms' efficiency, as large amounts of meaningful information are being omitted in this deluge of experimental results. The need for transforming large amounts of data obtained from the life sciences drives the development of statistical and data mining algorithms for the fusion and validation of biomedical experiments.

Nowadays, there still remains a great amount of knowledge to discover regarding the molecular mechanisms underlying disease. This information is vital, especially in the case of applications for constantly developing personalized medicine. The customization of therapies is a pressing issue when considering the numerous cases of diseases of affluence.

Heart and circulatory system diseases, cancer, diabetes have vastly increased in prevalence along with growing wealth in highly developed societies. However, these very diseases are now also becoming the leading killers in the developing world. This is why research in the field of diagnostics, prognostics and treatment is the key to elevating life expectancy and comfort around the globe.

Many of the leading mortality processes are rooted in different omics factors. Moreover, a shift can be observed in the ongoing studies, where instead of looking into single traits, combining information from different systems and their interactions is believed to hold the explanation to many unanswered questions in contemporary medicine.

Currently, statistical design of experiments allows for planning of complex studies while maintaining control over technical bias. The equal importance of performing tailored data processing in order to enhance quality of the results has been widely demonstrated. Furthermore, it has been previously shown that drawing attention towards effective and apt statistical analysis techniques and literature research is worthwhile, as it produces meaningful original biological conclusions.

Although the above discussed issues have been previously commonly acknowledged in the scientific community, the constantly changing database in terms of quantity and quality requires continuous efforts towards the optimization of data analysis pipelines. Therefore, this work has been dedicated to the investigation and implementation of comprehensive procedures enabling the integration of multi-omics data sets for discovery of disease biomarkers and their interactions. Different stages of data analysis have been covered from attentive preprocessing, which handles important sources of bias in high-throughput biological experiments, to the establishment of novel algorithms for combined analysis of data derived from multi-platform and multi-domain experiments.

1.2 Aim of the work

The goal of this work was to investigate diverse approaches for high-throughput molecular biology integrative data analysis to enable the discovery of disease of affluence biomarkers. The research methodology comprises a thorough overview of existing approaches for data combination, merging, comparison, and joint analysis, as well as the development of new methods for handling multi-omics studies. The expected outcomes of this work include the establishment of novel tools and procedures tailored to the tasks of multi-platform and multi-omics data and result integration.

Based on the motivation and aim of this thesis, the following statements have been formulated:

1. Adequate preprocessing of high-throughput molecular biology data, including identification and correction of batch effects allows to avoid discarding valuable potential discoveries.

2. The introduction of customized approaches for integrative analysis of data sets acquired in twin experiments within an omics allows for the improvement of statistical inference and classification tasks.

3. Statistical integration of multi-omics and different cell system data constitutes a means for single data set validation and leads to comprehensive mechanism characterization through the contribution of novel meaningful biological conclusions.

1.3 Chapter contents

The Background chapter contains an introduction to the problem of biomarker discovery and diseases of affluence. Furthermore, current high-throughput molecular biology techniques are presented, and lastly, various aspects of heretofore applied biomedical data integration have been described.

The Materials and Methods chapter presents the analysis methods introduced in this work and provides a description of the data sets utilized for implementation and testing. Firstly, the proposed Batch effect Identification (BatchI) using dynamic programming algorithm is explained. Next, a multi-platform transcriptomics data integration pipeline using statistical integration and dose profile based preselection for classification is discussed in detail. Futhermore, an inter-omics data integration approach is shown on a transcriptomics and proteomics data set. Finally, tools used for different tissue data integration are presented.

The Results and Discussion chapter presents the most important findings derived from the presented analyses. The BatchI algorithm is tested on multiple data sets with known in advance and unknown batch structure. The inter-platform transcriptomics analyses are discussed in terms of differential expression identification and emerged dose response biomarkers for their classification utility. The multi-omics approach is presented as an advantageous technique for single-omics experiment validation. The inter-tissue analysis workflow is considered as a tool for significant factor identification in sets of diverse cell systems data.

The Conclusions chapter summarizes the most significant achievements accomplished in the course of the work presented in this dissertation.

Chapter 2

Background

2.1 Biomarker discovery

In medical terms, a biomarker is an objective characteristic indicating the occurrence and/or severity of a disease. The value of such signals is inestimable, as in case of nearly any disease diagnosis time plays a key role in the application of appropriate therapy and is essential for increasing life expectancy and comfort. Therefore, immense effort is being put into scientific discovery of efficient disease biomarkers. A variety of biomarkers is already used in clinical tests, among others: blood, urine, lymphocyte, tooth enamel tests. The easier and faster a sample collection method is for a certain biomarker, the more attractive it is as a screening method.

Especially promising biomarker discovery techniques come from the field of molecular biology. They have great potential due to the possibility of investigating multiple aspects at a time, however, they possess also technical limitations. The current methods may be grouped into five "omics" domains:

- Genomics - derived from studying the genome by means of e.g. gene expression or sequence measurements
- Proteomics - relying on the identification of proteins in a sample and their levels
- Metabolomics - referring to global analysis of the set of metabolites
- Lipidomics - addressing the analysis of lipids through spectrometry and chromatography techniques
- Glycomics - studying the commonly occurring post-translational protein modifications.

The "omics" neologism addresses fields of study which have the objective of comprehensively characterizing biological molecules that translate into the structure, function, and dynamics of an organism. These methods more often than not, rely on high-throughput techniques for sample processing and obtaining data.

In most medical conditions (e.g. cancer), biomarkers may be classified into usage categories: predictive, prognostic and diagnostic. The first group is used with the aim of predicting response to treatment, the second group participates in estimating the risk associated with disease progression, and the third group serves as indicators of illness (Goossens et al., 2015).

Biomarker discovery, despite its great potential, is a challenging task due to multiple issues resulting from the different development stages. These problems may concern clear definition of research questions and experimental design, assay reproducibility, sample costs and availability, legislation and infrastructure obstacles. All these factors contribute to a low rate of successfully implemented biomarkers in the clinical setting. The first step of biomarker development, namely identification, is primarily executed using one of the two approaches: statistical or knowledge based. It has been now recognized that efficient tools for enabling the combined use of these two approaches are the key to successful biomarker transferring to the clinic (McDermott et al., 2013).

2.2 Diseases of affluence and radiation

In the case of many contemporary diseases their occurrence is most common in the higher developed regions of the world. Hence, a certain group has been given the name: diseases of affluence (Howe and Loraine, 2013). The focus on this group is strong due to its comprising of some of the most considerable health hazards in our societies:

- obesity
- cardiovascular diseases
- some cancers (mainly colorectal)
- type 2 diabetes
- gout

- depression
- diseases related with vitamin and mineral deficiency.

The causes of these diseases are being associated with modern lifestyle, including dietary habits and sedentary daily routines. Alas, more importantly, the high prevalence of these diseases makes it a vital task to develop knowledge about the mechanisms and, subsequently, produce solutions for curing the vast population of people affected.

Of the above, cancer and cardiovascular disease are the leading causes of mortality. When considering the two, radiation often plays a key role in both. In the former, radiotherapy is currently a substantial part of the treatment process, used in a majority of cases, whereas in the latter, exposure to radiation is a major incidence factor. Ionizing radiation is an omnipresent factor, which has a significant impact on many aspects of human life. Small doses are absorbed on an everyday basis while using everyday equipment, and higher doses occurring during accidents may have extremely detrimental effects (Abbott et al., 2015). Moreover, medical procedures such as radiation therapy constitute the leading cause of man-made ionizing radiation (Ray et al., 2012).

Radiotherapy consists of cancer treatment and pain reduction by means of ionizing radiation. Some of the most often types of cancer treated this way include: breast, lung, cervix, prostate, head and neck. Despite well established medical procedures for the use of radiotherapy, many patients suffer from adverse effects due to radiation toxicity. These may reduce life quality drastically and include hair loss, diarrhea, nausea, changes in the urinary and reproductive systems, metastases, lymphedema, arthritis. This response to treatment is conditioned by radiosensitivity, which is an individual factor indicating a person's susceptibility to harmful effects of radiation exposure. It is estimated that treating these adverse effects exceeds the costs of radiotherapy itself. Moreover, radiation doses applied in cases of lung and breast cancer increase the chance of developing heart disease by ca. 50%. The personalization in terms of dose application frequency and quantity, considering whether the patient is radiosensitive or radioresistant, would greatly help therapy planning and prognostics. One of the potential effective biomarkers of this trait are lymphocytes due to their high radiosensitivity and facility of sample collection.

Radiation-induced ischemic heart disease occurs when the blood vessels are subjected to radiation and oxidative stress activates inflammatory response leading to the formation of foam cells blocking free blood flow and acting pathogenically (Taunk et al., 2015). There is growing evidence that the pathogenesis of IHD and cancer shares common pathways and preventive strategies (Masoudkabir et al., 2017).

2.3 High-throughput molecular biology techniques

High-throughput screening is a branch of experimentation methods conducted on a large scale by parallel investigation of thousands of features. In the biomedical applications it is particularly applicable to drug and biomarker discovery. These types of analyses may be conducted on many levels of compounds in terms of the omics they represent. Bioinformatics has from the start accompanied high-throughput techniques and was necessary to enable robust analysis. The experiments may be conducted on different levels: DNA (genomics), RNA (transcriptomics), protein (proteomics), metabolites, etc. The first molecular biology experiments described as high-throughput included microarrays for genomics and transcriptomics measurements, and mass spectrometry for proteomics (Baggerly et al., 2006). Soon however, in terms of genomic and transcriptomic studies massive parallel sequencing became the top technology for investigating gene variants and expression (Widłak, 2013).

The techniques currently in use as high-throughput molecular biology are:

- genomics - next-generation DNA sequencing and microarrays
- transcriptomics - RNA-sequencing and microarrays
- proteomics, lipidomics, metabolomics - mass spectrometry

The main principles of these techniques are focused around the central dogma of molecular biology (Figure 2.1). This rule states that gene expression occurs through transcription of information from DNA to RNA, and then translation of RNA to respective amino acids forming proteins. The aforementioned techniques assess the quality and quantity of this process by experimental insight into molecules on one or more stages of gene expression, and the accompanying mechanisms.

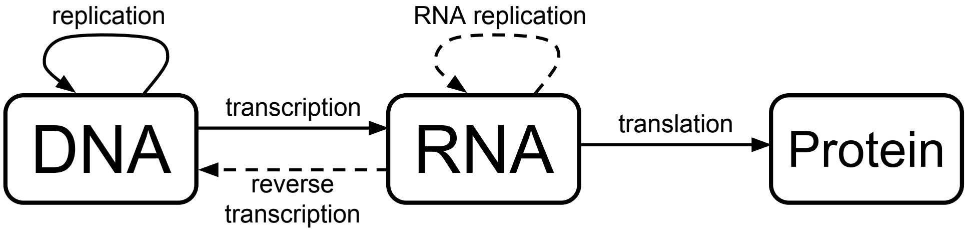


Figure 2.1: Illustration of the central dogma of molecular biology. Solid lines represent the general direction of gene expression. Dashed lines correspond to special transfers of biological sequential information

Microarray technology allows for the measurement of expression levels in thousands of genes at a time (Govindarajan et al., 2012). It is based on the application of multiple spots of DNA fragments attached to a solid plate and used to assess the quantity of RNA present in a sample. This method is based on the process of hybridization - the property complementary nucleotides have to specifically pair with each other. DNA fragments from one strand are present on the microarray chip, while the assessed sample of mRNA or cDNA is shredded, the fragments amplified and applied to the chip for hybridization. Once a DNA fragment hybridizes at a specific spot, a fluorescent labeling substance is released to be caught by sensors in the scanning process. As the fragments present on a microarray are specific to particular genes, they are effective means of quantifying the corresponding gene expression (Figure 2.2). Moreover, this high specificity enables using microarray technology for the purpose of single nucleotide polymorphism (SNP) detection (Heller, 2002).

Next-generation sequencing (NGS) technology was developed for the purpose of SNP detection and gene expression measurements not only for known sequences, but also for high-throughput processing of millions of DNA fragments for gene discovery. DNA sequencing is used for the determination of nucleotide order in a molecule. This was first possible with the use of Sanger sequencing, which served as the basic technique for carrying out the Human Genome Project. The underlying principles formed what is now known as massively parallel sequencing (also called next-generation sequencing or high-throughput sequencing: HTS). The biological mechanisms in this

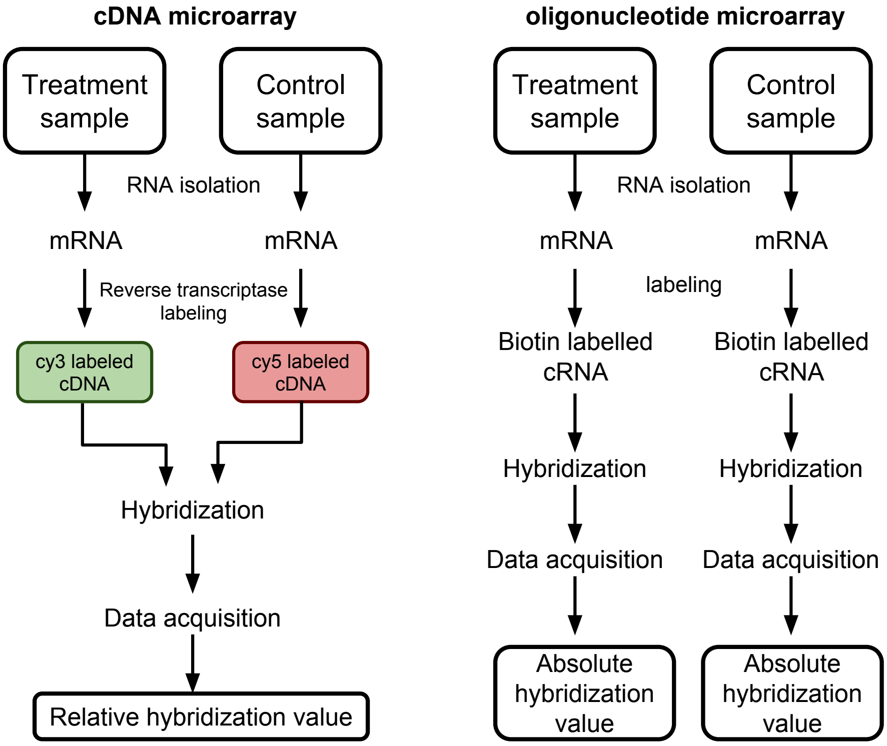


Figure 2.2: Microarray data processing workflows. The left diagram shows data processing using two-channel cDNA microarrays, where the raw data are a ratio of the hybridization intensity between treatment and control samples. The right diagram underlines the difference in oligonucleotide microarrays, where treatment and control samples are measured independently.

method consist of fragmenting the genomic strand and identifying the subsequent nucleotides based on signals emitted while ligating to a template strand (Figure 2.3). Traditional Sanger sequencing required performing all the necessary steps one by one: sequencing, separation and data acquisition. NGS, relying on array-based sequencing, introduced a strong efficiency improvement, as it allows for the combination of all of the previously sequential methods into millions of parallel processes (Mardis, 2008). Nowadays, sequencing the human genome is available within a few hours, as opposed to the Human Genome Project which lasted 13 years (Venter et al., 2001).

Mass spectrometry (MS) is a technique used in proteomics for high-throughput determination of protein and cellular functions. It is an important tool specifically for primary protein sequence analyses, post-translational modifications and protein-protein interactions (Aebersold and Mann, 2003). The mechanisms behind this method lie in ionizing the molecules in the gas state and measuring their mass-to-charge ratio: [m/z] (Figure 2.4). For this purpose, mass spectrometers, regardless of their technological

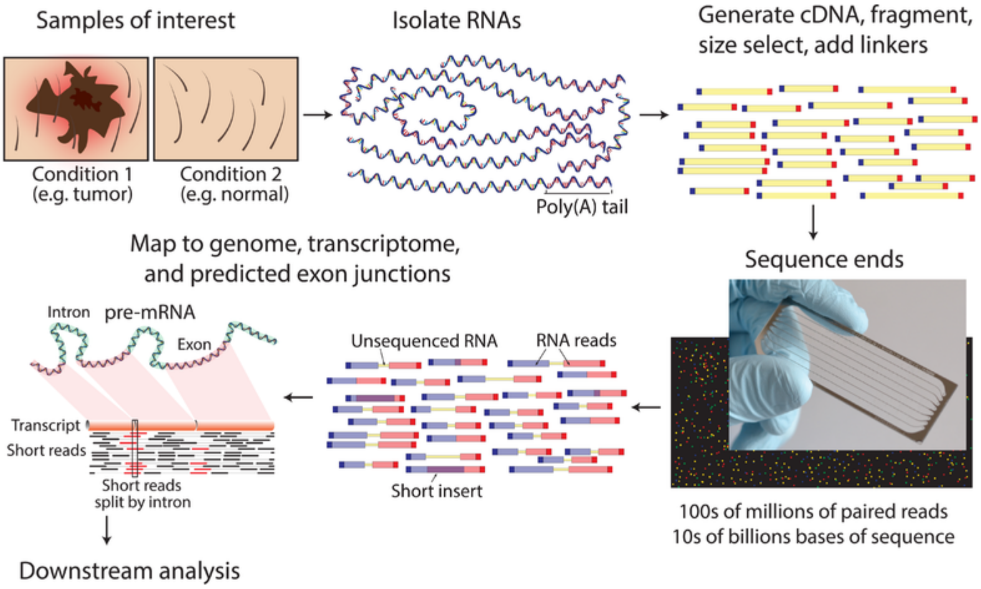

Samples of interest
Isolate RNAs
Generate cDNA, fragment, size select, add linkers
Condition 1 (e.g. tumor)
Condition 2 (e.g. normal)
Poly(A) tail
Map to genome, transcriptome, and predicted exon junctions
Sequence ends
Intron
pre-mRNA
Exon
Unsequenced RNA
RNA reads
Transcript
Short reads
Short reads split by intron
Short insert
100s of millions of paired reads
10s of billions bases of sequence
Downstream analysis

Figure 2.3: Next generation sequencing analysis pipeline. (https://commons.wikimedia.org)

differences, are all based on three components: an ion source, an analyzer for separating the ionized particles, and a detector (Han et al., 2008). Among the most common applications of MS technology are protein identification and quantification. The latter makes it useful for biomarker identification, as it enables the detection of different levels of protein between samples of diverse characteristics (Rifai et al., 2006). Often this procedure is coupled with liquid chromatography for initial separation of the analyzed fractions.

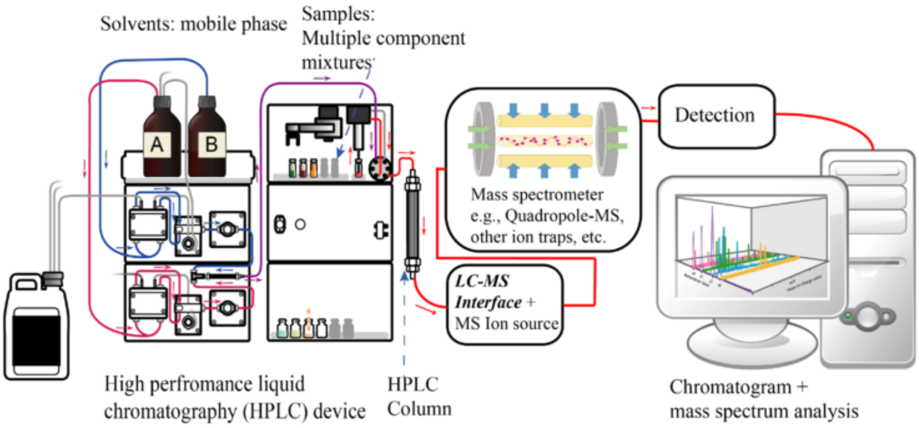

Solvents: mobile phase
Samples: Multiple component mixtures
A
B
Detection
Mass spectrometer e.g., Quadropole-MS, other ion traps, etc.
LC-MS Interface + MS Ion source
High perfromance liquid chromatography (HPLC) device
HPLC Column
Chromatogram + mass spectrum analysis

Figure 2.4: Liquid chromatography coupled with mass spectrometry experiment workflow. (https://commons.wikimedia.org)

2.4 Integration in omics data

Multi-omics data integration became a natural continuation of analysis techniques in many fields of molecular biology. In order to obtain a comprehensive explanation of studied phenomena in the omics domains, merging and processing of data from different experiments have become indispensable in the analysis workflow. In general, the notion refers to combining data residing in different sources and providing users with a unified view of these data (Lenzerini, 2002). The multitude of experimental techniques and variety of statistics, data mining and machine learning tools developed over the years, provided a plethora of means and paths to interpret the term: **integration**.

2.4.1 Integration within an experiment

Analyzing data from a single experiment already requires a careful and accurate choice of techniques for data preprocessing. Although, it is often not perceived as such, the normalization and standardization of data is in fact a step towards integration of data from samples collected and processed within an experiment. High-throughput techniques are especially prone to technical bias due to the usual large scale of an experiment, and therefore, it is particularly essential to select appropriate preprocessing methods.

Normalization and standardization methods vary in different experiments. In microarrays there are several algorithms available for normalizing data from individual hybridizations. However, most of them comprise the following steps: background adjustment, data normalization, and in the case of oligonuclotide arrays, where probe copies are scattered throughout the chip, a summarization step (Quackenbush, 2002). For sequencing data, specifically in RNA-seq where read counts serve as estimates of gene expression levels, multiple measures and techniques have been proposed for normalization when taking into account the position in the genome, gene length and the overall count distribution (Li et al., 2015). When dealing with mass spectrometry data, bias caused by instrumentation is not to be overlooked, and therefore, intra- as well as intergroup normalization is necessary. The methods proven most effective and common are those based on variance stabilization (Välikangas et al., 2016).

Although there exists a multitude of techniques for data normalization, depending on the type of experiment, one preprocessing step should be universally carried out, regardless of the high-throughput technique, i.e. batch effect filtration. Batch effects are technical sources of variation, separating samples into subgroups according to their quality traits instead of the biological or scientific studied condition, seen in a wide range of high-dimensional molecular biology experiments (Scherer, 2009). The factors contributing to batch effect occurrence are e.g. differences in sample processing protocols, different experimentalists, or changes in external conditions prevailing during data acquisition. These systematic errors may be understood as batches of samples processed together in an experiment. This means that the size of a batch is defined by the capacity of a machine (Figure 2.5). Other common sources of batch effects are uncontrollable changes of some/many of the experimental conditions over time (Leek et al., 2010). In high-throughput experiments batch effect bias is unavoidable, occurs with different experimental platforms, survives standard normalization and correction procedures and leads to significant errors in data analyses, like the decrease of sensitivity or increased number of false discoveries (Chen et al., 2011; Luo et al., 2010). It has been demonstrated by numerous studies that identification and correction of batch effects can substantially improve results of data analyses (Sun et al., 2011; Auer and Doerge, 2010; Sims et al., 2008).

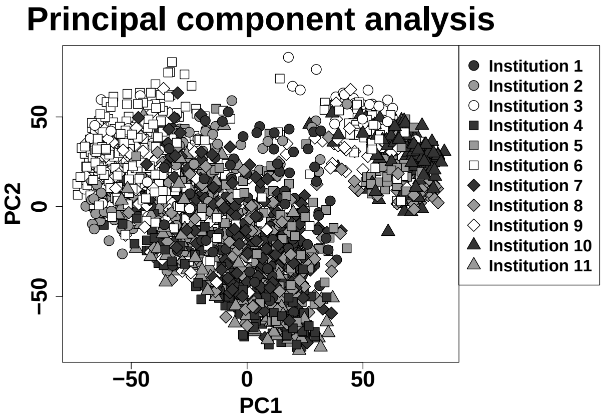


Figure 2.5: Illustration of batch effect on the example of the MILE study. The principal component analysis plot indicated batch effect existing due to samples being processed in different institutions, despite efforts made to retain identical experimental protocols (Labaj et al., 2017).

It is therefore of primary importance that batch effect should be recognized and filtered from data sets. Research results than have been compromised by the lack of batch effect management provoked the development of a variety of batch effect correction algorithms. The issue was first observed over the course of microarray experiments, and therefore many of these techniques have been developed for microarray data, however, since then adjustments have been made and new proposals contributed for the purpose of multi-omics data processing. The first attempts relied on mean centering or were ratio-based, yet the need for more sophisticated approaches arose promptly. (Benito et al., 2004) developed a method called distance-weighted discrimination, based on support vector machines (SVM) classification algorithm for detecting and removing batch biases. SVM algorithm is used for computing a separating hyperplane between data points corresponding to different batches. Then, the obtained parameters are used to remove batch bias. (Bylesjö et al., 2007) use a multivariate regression model with hidden elements, called orthogonal projections to latent structures (Trygg and Wold, 2002) for identification and correction of batch biases. The case of gene expression data in microarray experiments enabled the creation of a family of RUV (Remove Unwanted Variation) methods, specifically for the purpose of handling these data, based on applying negative control genes for batch effect adjustment (Gagnon-Bartsch and Speed, 2012). This knowledge driven approach, however, limits the usability to a narrow group of experimental techniques where such negative control features are possible to describe. A method named ComBat (Combating Batch Effects When Combining Batches) for removing batch effects in DNA microarray data, based on the empirical Bayes approach, was proposed by (Johnson et al., 2007). They define and estimate additive and multiplicative batch bias parameters and then use them to modify distributions of gene expression. The approach was proven reliable, useful for data sets with multiple batches and robust to small sample sizes and may be extended to other experimental techniques (RNA-seq, genomics, proteomics).

The above mentioned approaches, generally rely strongly on the information about batch grouping structure. However, often it is not the case that these data are available considering the frequent lack of records concerning experimental conditions, its incompleteness, or a degree of ignorance towards factors which may influence batch effect occurrence. Thus, a need for the identification of batch partitioning has been perceived and diverse methods developed for the purpose of detecting existence of batch effects and estimating the proportion of variation in the data resulting from batch effects. (Alter et al., 2000) apply PCA to genome-wide expression data and propose removal of noisy components (eigengenes) corresponding to low singular values. Under the assumption that one (some) of the noisy eigengenes corresponds to batch effect the use of the method by Alter et al. leads to batch effect correction. (Reese et al., 2013) present an extension of PCA to quantify the existence of batch effects, called guided PCA (gPCA). They derived a test statistic, based on the traditional PCA and gPCA, for detecting batch effects. The test statistic, δ, quantifies the proportion of variance owing to batch effects. Surrogate variable analysis (SVA) (Leek and Storey, 2007) is an algorithm for combined batch effect identification and correction by means of effect estimation. (Yi et al., 2017) proposed another approach for hidden batch effect identification based on data-adaptive shrinkage, coupled with a regularization technique of non-negative matrix factorization for batch effect correction.

Applying filtration of batch effects is a significant step towards enabling the integration of data within an experiment performed at different times, conditions or laboratories. As mentioned, various methods have been established for data adjustment to account for suspected batch effect present in the data. However, the identification of unknown batch effects still remains as a subject for development. In this work an approach based on dynamic programming is proposed for identifying batches in data that may be sorted (on a timescale or otherwise).

2.4.2 Integration within an omics

High-throughput techniques in molecular biology are constantly being improved in many aspects. This naturally entails the competition of multiple scientific and technological centers in the challenge for developing the most effective platforms. This situation undoubtedly has main advantages in the advancement of modern science, however, it also implies certain issues in the subsequent data analysis workflows. The presence of various experimental platforms is tantamount with the existence of different standards and if data from numerous experiments is to be merged, appropriate measures in the analysis phases need to be taken.

In nearly any omics field, one will not find a gold standard technique, but rather a choice of well established experimental platforms. As such, when considering microarray platforms aside from the most popular Affymetrix oligonucleotide chips, a range of commercial arrays is in use (Agilent, Life Technologies, Qiagen, to name a few), as well as custom cDNA microarrays, which were initially more frequently encountered. On the other hand, with the genome sequencing industry on the rise currently, the commercially available sequencers (such as Illumina, Roche 454) are not only constantly being enhanced, but also innovative, previously unavailable solutions are being introduced to the market (Oxford Nanopore, Pacific Biosciences). In all these cases, in principle the same research may be conducted. Nonetheless, when it comes to performing a joint analysis of two data sets derived from different platforms, supplementary actions need to be taken in all stages of the analysis: from preprocessing, through downstream inference, to functional validation.

The validation of results from a single high-throughput experiment may take various forms when the biological context of the available information is the same. The most desirable way is through biological validation where an experimental technique is available to confirm the disclosed findings. For instance, in the case of microarray data analysis, where gene expression levels are assessed indirectly, qPCR providing a direct expression level measure would be an appropriate validation technique. However, these methods, while producing the most reliable results, are usually costly and time consuming. Thus, *in silico* validation procedures became an attractive alternative. For that purpose, it is possible to carry out functional analyses (using bioinformatics repositories, such as Gene Ontology or KEGG pathways), statistical validation by means of multiple testing correction, or result verification on an independent data set. This last method, depending on the available resources, may be performed through literature research or by means of analyzing data obtained it the course of a similar experiment. This potential lying in vast amounts of data from experiments makes it crucial to extract information entirely efficiently in single studies, but also to make the most of combining information from already available data and knowledge. The need for transforming large amounts of data coming from the life sciences drives the development of data mining algorithms for the fusion and validation of biomedical experiments. Combining information from available data sets is becoming an essential tool for scientific researchers.

In previous studies, various methodologies have been considered for combining biochip data sets across platforms. As simple approaches such as standardization and mean-centering had their limitations, more complex concepts started to emerge. (Parmigiani et al., 2002) introduced the Probability of Expression method, which transforms expression data to signed probabilities. (Breitling et al., 2004) present the Rank Product computation scheme, (Shabalin et al., 2008) developed cross-platform normalization (XPN) based on iterative k-means clustering. These algorithms have been evaluated in numerous studies on merging multiple microarray data sets (Sîrbu et al., 2010; Liu et al., 2013), yet it seems that the question of integration of platforms of different nature has not been attended to. Hence, this work contemplates an approach that addresses the particularly intricate issue of combining data sets from two types of microarrays: oligonucleotide and cDNA in an integrative transcriptomics data analysis.

2.4.3 Inter-omics integration

The convoluted interaction network between the subjects of different omics studies is largely an enigma still to be unraveled. Nevertheless, the current state of knowledge in life sciences allows for the joining of certain pieces of the puzzle. In molecular biology, knowledge about the central dogma (Figure 2.1) enables searching for common mechanisms on the level of genes, transcripts, proteins, and even further in areas such as metabolomics.

This richness of possibilities drives the development of algorithms and procedures for inter-omics data analyses. The complexity and multi-level character of the data network require in each individual case customized, tailored approaches on the border between statistical tools, machine learning techniques and big data analyses. Adaptive techniques are key to unfolding the mechanisms underlying disease and other biological conditions on a multilevel scale. When studying the response of genes to a certain stress factor it is often not straightforward to infer that the corresponding protein products will function accordingly. The response may be completely opposite, or depend on a cascade of signals joining the studied features on a genomic and/or proteomic level and become even more ambiguous to explain. This ramification imposes a major shift from examining single traits to producing more comprehensive and detailed descriptions of studied processes. For this purpose, the combination of multi-omics level data proves to be the correct solution. Recently, this has been recognized in a number of studies including high-throughput data. As such, more and more attempts at combining knowledge from different omics for cancer research are being made (Dimitrakopoulos et al., 2018), merging them with clinical data (Zhu et al., 2017) and improving and expanding the data integration toolkit (Huang et al., 2017; Tini et al., 2017). In this study, the importance of inter-omics data analyses is demonstrated with an example of an original statistical and data mining workflow for processing transcriptomics and proteomics data sets.

2.4.4 Inter-tissue integration

Experiments in the different omics fields are all an attempt to build foundations underlying knowledge concerning biological processes. However, shifting the scale to examining mechanisms occurring in entire cell systems is a no less important task to be addressed. Several experimental techniques have the capabilities to reveal tissue architecture, generating a wealth of biological knowledge and a better understanding of many diseases, especially with single-cell sequencing on the rise (Chen et al., 2018). Identifying regulatory elements from different cell types is necessary for understanding the mechanisms controlling cell type-specific and housekeeping gene expression (Xi et al., 2007; Xu et al., 2014). At the very beginning of microarray technology development it has been shown that expression patterns of diverse cell types contribute to the pathology (Heller et al., 1997). However, not only is the gene level suitable for inference on the issue of multiple cell systems operation, but also proteomics tools and methods contribute widely in the field. As such, efforts have been carried out in order to yield an inventory of the building blocks of the most commonly used systems in biological research. (Geiger et al., 2012) study eleven common cell lines to reveal high similarity in terms of expressed proteins, despite their distinct origins. Notably, the NIH Roadmap Epigenomics Consortium generated the largest collection of human epigenomes for primary cells and tissues (Kundaje et al., 2015).

The aforementioned studies do not fully respond to the challenging task of establishing a set of tools for integrative analysis of data acquired from different cell systems. In this work, techniques for examining the similarity between various cell systems are presented in application to a study on effects of irradiation on exosomes. Exosomes are specialized vesicles derived from endocytic compartments that are released by many cell types. Small RNA loading into exosomes and transfer to recipient cells plays a role in intercellular communication (Zomer et al., 2010). The main functions of exosomes include membrane exchange between cells, alternative to lysosomal degradation, transfer of antigens from tumor to dendritic cells (Edgar, 2016). The deciphering of mechanisms governing this communication under different biological conditions will lead to discoveries in the process of promoting tumor progression.

2.5 Data mining and statistical integration methods

The advantages of incorporating bioinformatics databases into biomarker discovery schemes have been previously shown in various studies (Meehan et al., 2013; Kong et al., 2014). Recently, much focus has been directed towards the development of methods, algorithms and procedures for multi-omics data integration, especially in order to broaden horizons in the field of precision medicine (Huang et al., 2017). The emerging results already introduced a significant impact in the diagnostics and prognostics of cancer and other diseases (Li et al., 2018; Bakker et al., 2018). Moreover, the incorporation of complex deep learning techniques into biomedicine analyses is starting to play a key role in cancer patient survival prediction (Chaudhary et al., 2018). Finally, newly developed workflows promise crucial advancements in biomedical research and beyond (Kohl et al., 2014; Gajula, 2016).

Apart from the dynamically progressing field of data mining and machine learning, this work presents the utility of sophisticated statistical analysis tools for multi-omics data integration. Upon recognizing that standard comparisons of multiple experimental results are often limited, due to testing with fixed thresholds, even when applying classical correction methods for multiple testing such as false discovery rate (FDR), this study is partly devoted to the applicability of statistical data integration methods. Although in some cases, it may be possible after careful normalization and batch effect correction to combine several data sets into one, in this work the possibility of statistical testing p-value integration is explored. The approach enables merging data sets regardless of the original omics field and experimental nature, provided that the studied features are analogous among the single sets. It is an intermediate method between merging data at the initial stage and combined analysis of only the final results.

P-value combination is a statistical concept that was first introduced by (Fisher, 2006). It is based on the assumption that the p-values come from tests on independent experiments and the resulting combined p-value is derived from a distribution of log-transformed average p-values. From then on, the method was developed, extended and modified multiple times, in relation to diverse data sets and requirements. The principle between all of the proposed methods is similar and may be illustrated with the graph in Figure 2.6.

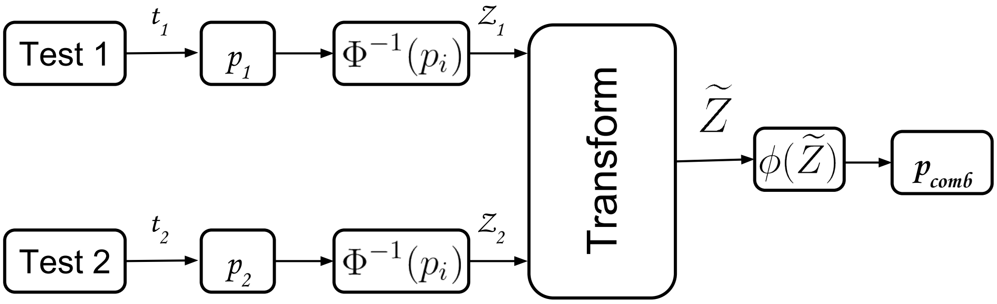


Figure 2.6: General p-value combination procedure.

The individual p-values are transformed into statistics based on a given distribution, then they are merged and the resulting p-value is calculated based on the combined new overall statistic. The main approaches used for p-value combination are

Lancaster's modification of Fisher's p-value method (Lancaster, 1961), Stouffer's design for symmetrical distributions (Stouffer et al., 1949), and diverse weighted Z-score methods (Liptak, 1958). The choice of method for data set integration depends on the character, balance and distribution within the particular data sets.

Chapter 3

Materials and Methods

3.1 Batch effect identification using dynamic programming

Batch effect correction tools enable filtration of confounding factors from data sets and in this way enhance analysis results by driving the main focus towards biological variability. However, if for instance laboratories, where different samples are processed, are considered sources of batch effects, correcting for these effects becomes part of the data integration pipeline within an experiment. In many experiments, information about batch structure is not provided though and state-of-the-art procedures often depend upon this information. Therefore, in the course of this work a novel batch effect identification algorithm (Papiez et al., 2018b) has been proposed and tested on a number of experimental data sets. These include series of DNA microarray, mass spectrometry (MS) and RNA-seq measurements.

The dynamic programming identification procedure requires the representation of each sample with a quality index (QI). In the microarray experiments it is defined by the average intensities among all features. For the MS data the Total Ion Current (TIC) for each sample is applied and for the RNA-seq data the median number of counts. The quality index may be also any chosen statistic representing the data levels in a single sample. It is worth underlining, that since the objective in batch effect handling is to account for sources of technical variation, it is advisable to calculate the summarizing quality index on data at as early a stage of processing as possible.

In this sense, the issue of batch identification may be defined as dividing a sorted series of samples into a number of batches, such that a sum of absolute deviations of the quality indexes within a batch is minimized. This task is accomplished by partitioning the range of quality indexes of samples into bins (batches) using the dynamic programming algorithm (Bellman, 1961; Jackson et al., 2005).

Indexes of samples in the experiment are denoted $i = 1, 2, \dots, N$. The division into subgroups involves defining K batches, $B_1, B_2, \dots B_K$, where the $k - th$ batch is the range of indexes $B_k = B(i, i+1, \dots, j) = i, i+1, \dots, j$. The quality index is denoted by QI_i. Absolute deviation of the QI within batch B_k is:

$$AbsDev(B_k) = \sum_{l \in B_k} |QI_l - \overline{QI}_{B_k}| \tag{3.1}$$

The minimization index for the dynamic programming algorithm is the sum of absolute deviations

$$I(K) = \sum_{k=1}^{K} AbsDev(B_k) \tag{3.2}$$

Optimal partitioning $B_1^{opt}, B_2^{opt}, \dots B_K^{opt}$ leads to a minimal value of the sum of absolute deviation indexes corresponding to all batches:

$$I_{1 \dots N}^{opt}(K) = min_{partitions}^{1 \dots N}[\sum_{k=1}^{K} AbsDev(B_k)] \tag{3.3}$$

The upper index of the above minimization operator, $1 \dots N$, represents the range of time indexes of samples, while the lower one indicates that minimization is over all possible partitions. In order to formulate dynamic programming recursion an optimal partial cumulative index for the range of samples $1, 2, \dots, j$ is calculated:

$$OCI_{1 \dots j}(k) = min_{partitions}^{1 \dots j}[\sum_{\chi=1}^{K} AbsDev(B_\chi)] \tag{3.4}$$

Dynamic programming recursive procedure, called Bellman equation, can be written in the following form:

$$OCI_{1...j}(k+1) = min_{i=1...j-1}[OCI_{1...i-1}(k) + AbsDev(B(i, i+1, \ldots, j))] \quad (3.5)$$

Iteration of the above Bellman equation provides the retrieval of the optimal partition $B_1^{opt}, B_2^{opt}, \ldots B_K^{opt}$ and the optimal (minimal) value of the sum of absolute deviations index $I_{1...N}^{opt}(K)$. The algorithm will not allow the condition that one batch contains fewer than three samples, as a smaller number would be insufficient to calculate dispersion metrics. The analysis relies on computing variance related statistics in consecutive analysis stages. The implementation is also designed in such a way, that the parameter is modifiable to set the minimum threshold to a number larger than three.

3.1.1 Batch number selection

The proposed method includes a parameter that requires setting, namely the number of batches into which the data should be divided. This may be executed by dividing data into a number of batches from 1 to K and in each of these partitioned sets calculating the δ gPCA statistic as described in (Reese et al., 2013), which is defined as the proportion of total variance due to batch and may be calculated as a ratio of variance of the first principal component in guided PCA (taking into account batch effects) to variance of the first principal component in unguided PCA.

$$\delta = \frac{var(XV_{g1})}{var(XV_{u1})} \quad (3.6)$$

In order to estimate the δ statistic sampling distribution, M permuted data sets are generated by randomly shuffling the partitioning of samples to batches. Then, for each assignment calculation of δ_{PERM} permuted gPCA statistic is performed. The position of the actual test statistic δ among the generated δ_{PERM} test statistics gives an adequate p-value, which may be described by defining if δ is significantly greater than would be obtained by chance. This approach also accounts for a situation where batch effect partitioning is not necessary. If the statistic cannot be deemed significant, it means that batch effect is negligible, and identification and correction are irrelevant. In any other case, batch assignment corresponding to the lowest p-value is fixed as the optimal number of batches.

During the testing phase, when setting the value of M to 1000 (default value set in (Reese et al., 2013)), in none of the experiments was a value δ_{PERM} greater than δ reached. This renders choosing the optimal number of batches in a data set not possible. As the computation time is proportionate to the number of permutations M, increasing the number by an order of magnitude raises this time dramatically. What is more, the δ statistic distribution differs in every data set and may adopt multimodal shapes. In order to mitigate this issue, the use of a kernel density estimator is proposed, which provides plausible approximations of the δ statistic distribution. When considering a permuted gPCA statistic δ_{PERM}, the underlying probability density function f used to generate this sample can be approximated using the kernel density estimator given by:

$$\hat{f}(\delta) = \frac{1}{K}\sum_{i=1}^{k} kernel(\delta, \delta_i) \tag{3.7}$$

where $kernel$ is a kernel function. For the purpose of this application $kernel$ is chosen as a standard Gaussian function:

$$kernel(\delta, \delta_i) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(\delta-\delta_i)^2}{2h^2}} \tag{3.8}$$

where h is the bandwidth that controls the degree of smoothness of $\hat{f}(\delta)$. If h is chosen too small, the resulting estimate is usually overfitted with regard to the available samples. On the contrary, if h is too large, the resulting density becomes oversmoothed, with a simultaneous reduction of its variance across different samples. To select the bandwidth parameter a rule-of-thumb method available in the R stats package (Silverman, 1986) is used. The final p-value is obtained by calculating the area under the estimated distribution in the right tail from the observed δ statistic value.

The batch effect identification algorithm has been implemented in the BatchI R package and is available for download and use (Papiez et al., 2018b).

3.1.2 Data

The dynamic programming based method for batch identification was evaluated on a number of microarray and RNA-seq data sets obtained through the ArrayExpress (Kolesnikov et al., 2014) repository and an MS data set acquired in collaboration the Center of Oncology - Maria Sklodowska-Curie Memorial Institute in Gliwice.

The first step was to test the algorithm on data with known *a priori* batch structure. For this purpose, two sets of microarray data were investigated, E-GEOD-19419 (Walter et al., 2010) which consisted of gene expression profiles from peripheral blood of patients affected by neurological movement disorder DYT1 dystonia, containing 60 samples: 15 controls, 23 symptomatic and 22 carriers. The other one was E-GEOD-36398 (Rahimov et al., 2012) comprising gene expression profiles of tissues from two different muscles in patients with facioscapulohumeral muscular dystrophy and their unaffected first degree relatives, containing 50 samples: 24 controls and 26 FSHD. These experiments were carried out on HuGene 1.0 ST microarrays with 32321 measured probes. In both of the data sets samples were assigned to batches due to the differences in time of sample preparation and experiment performance. They include, respectively, three (E-GEOD-19419) and five batches (E-GEOD-36398).

The E-GEOD-65683 RNA-seq measurement data set was acquired from an experiment where sperm from male partners of couples undergoing fertility treatment was assessed. The study consisted of 72 samples split into 3 groups: 7 in group I, 56 in group II, and 9 in group III. The metadata included dates of the sequencing run performances, which served as information to divide the data into three batches.

The MS data was collected in a study investigating pulmonary cancer among smokers. In this case, an unfortunate design of experiment lead to the samples being processed in three distinct batches according to date (Pietrowska et al., 2012). The data consists of 377 samples: 282 controls and 95 cancer cases and a total of 700 protein features was detected.

Secondly, studies without the information about batch assignment were used to validate the method. Experiments E-GEOD-2034, E-GEOD-4183 and E-GEOD-10927, have been chosen for the analysis on the basis of being described as demonstrating a high proportion of variance due to batch effects in (Parker et al., 2014). The study E-GEOD-2034 (Wang et al., 2005) attempts at predicting the occurrence of distant metastases in patients suffering from lymph-node-negative primary breast cancer. The gene expression profiles were obtained from frozen tumor samples. The experiment labeled E-GEOD-4183 (Galamb et al., 2008) comprises gene expression profiles measured in colon biopsy samples using high-density oligonucleotide microarrays for the purpose of predicting local pathophysiological alterations and functional classification of adenoma, colorectal carcinomas and inflammatory bowel diseases. The last data set E-GEOD-10927 (Giordano et al., 2009) was acquired in a clinical study on molecular classification and prognostication of adrenocortical tumors by gene expression profiling.

For the gene expression microarray data sets the RMA normalization algorithm (Irizarry et al., 2003) was used for preprocessing. The RNA-seq data was aligned and processed for read counts using STAR (Dobin et al., 2013). The MS data samples were analyzed on a MALDI-ToF mass spectrometer in the mass range between 1,000 and 14,000 Da. Data preprocessing consisted of outlier spectra detection, global linear alignment, baseline correction, normalization and spectra alignment (Pietrowska et al., 2012). The identification of peptide ions in the spectra and the computation their relative abundances was achieved using a Gaussian mixture model based algorithm (Polanski et al., 2015).

3.1.3 GO Information Content functional analysis

Due to the unlabeled data sets E-GEOD-2034, E-GEOD-4183 and E-GEOD-10927 not having a reference partitioning into batches, the performance of BatchI algorithm was assessed by summarizing the relevance of biological findings revealed after the data was adjusted for batch effects. For this purpose, the Gene Ontology database was chosen. Gene Ontology is a comprehensive resource containing computable knowledge regarding the functions of genes and gene products. GO terms are structured in the form of a directed acyclic graph to provide information about the relationships between biological functions. The term names supply biological knowledge about the studied processes themselves, however, the position of an enriched GO term can be summarized with the Information Content measure.

Information Content (IC) is calculated as:

$$IC(t_i) = -ln(p(t_i)) \tag{3.9}$$

where $p(t_i)$ is the relative frequency of a term's occurrence, and may be expressed as a ratio of the probability of a term occurring in the corpus and the corpus being the set of annotations for all genes under consideration (Mistry and Pavlidis, 2008). This translates to the general rule where a term, which has a higher Information Content measure, is also more meaningful in the biological sense. The reason for this being that the further a GO term lies from the root node, the more specific the information is. Moreover, the fewer genes constitute a given ontology term, the more significant the overrepresentation becomes, because it is less likely to find genes linked to "smaller" terms by chance.

Hence, for comparative purposes, enriched gene ontologies were summarized for two scenarios: data analyzed **with and without** batch effect identification and correction. The overrepresented terms in each experiment were represented by the total sum of IC and standardized by dividing each GO term measure by its appropriate gene number size.

3.2 Multi-platform transcriptomics data integration

The ability to integrate and render data to be analyzed with reduced bias within an experiment is a pressing issue, however, it is only merging data across experiments that unlocks the full potential of integration methods. In this work two transcriptomics microrray data sets obtained using different platforms are analyzed in various aspects to present the proposed methods for merging data in order to enhance the information available from single analysis workflows.

3.2.1 Data Sets

The expression sets used in this study were obtained in the course of two independent microarray experiments on the subject of radiosensitivity. The experiments were designed with the objective of identifying genes differentiating radioresistant (RR) and radiosensitive (RS) women in a group of breast cancer patients undergoing radiotherapy. The clinical description of the samples and radiosensitivity status assignment is described in (Yarnold et al., 2005).

One experiment provided blood samples from 60 patients, of which 30 were classified as radiosensitive and 30 as radioresistant. This experiment was performed on the HuGene 1.0 ST Affymetrix oligonucleotide chips, measuring 19,718 genes, providing raw intensity CEL files. As for the second experiment, samples were gathered from 59 patients: 31 radiosensitive and 28 radioresistant (Finnon et al., 2012). It was carried out using a custom Breakthrough 20K cDNA microarray chip, measuring 19,959 genes, producing a set of GPR files produced by the GenePix 5.1 scanning software. The procedure was performed in a dye-swap manner, such that each sample was labeled with the cy3 and cy5 dye and hybridized to the chip against a reference sample from a pooled set of 30 breast cancer cell lines.

In both cases blood samples were collected from the donors for RNA extraction after 24h from lymphocytes for the amplification and labeling in the microarray experiment. The samples were divided into two lots labeled as one of the two conditions: controls and irradiated. In the oligonucleotide array experiment, one sample per patient was left as control, the other was irradiated with a therapeutic level dose of 2 Gy of X-rays. In the cDNA microarray experiment, the irradiated samples were subjected to a high dose of 4 Gy (Figure 3.1).

3.2.2 Preprocessing

The Affymetrix oligonucleotide single channel data was normalized using the Robust Multichip Average (RMA) method (Bolstad et al., 2003), which includes background intensity correction, quantile normalization and summarization using the median polish algorithm. Probes were reannotated with a custom chip description file [1] from the

[1] hugene10st_Hs_ENTREZG version 1.36.0 May 10, 2013

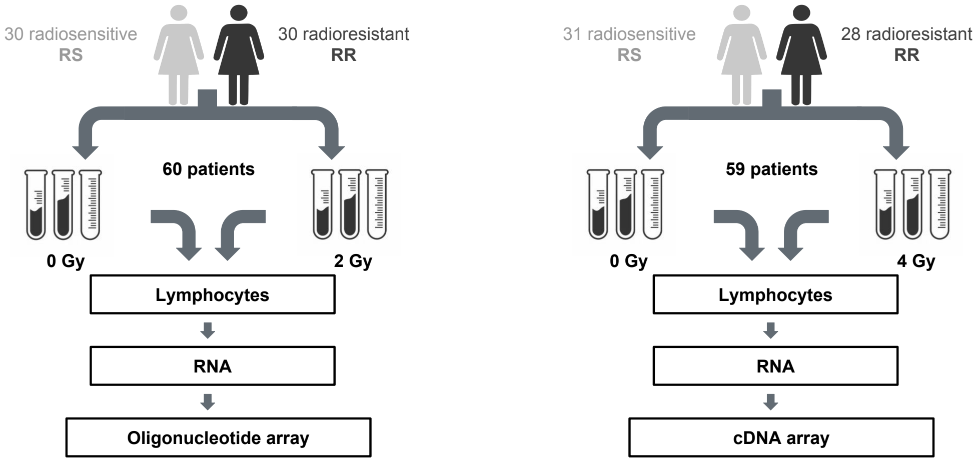


Figure 3.1: Diagram presenting a comparison of experimental designs. The twin experiments were carried out using the same labeling of RR and RS patients with similar numbers of samples. They differ nonetheless, with sample treatment doses and microarray experimental platforms. These issues had to be resolved during combined data processing.

Brainarray database (Dai et al., 2005).

The cDNA microarrays were preprocessed with the Bioconductor Limma package (Smyth, 2005). The values were background adjusted using the *normexp* algorithm. In order to retain compatibility between data coming from two platforms, the within array normalization step was omitted, as there is no equivalent in the oligonucleotide preprocessing pipeline, and between array normalization was executed with the *quantile* method. This resulted in an expression set of two replicates of patients' samples, one for each color channel (cy3 and cy5).

For the sake of comparison of data from two different microarray platforms, an approach was adopted where the main concern was to obtain data within the same space, in mathematical as well as biological terms (Papiez et al., 2014). Therefore, intensity data for separate color channels for the patients' samples was extracted and included for further investigation, excluding the information on breast cancer cell lines. This was motivated by the necessity of retaining consistency in terms of the biological representation of the signal, as there is no such reference available in the oligonucleotide chip experiment. Another reason was the lack of feasibility of juxtaposing expression values in oligonucleotide microarray data with ratios of expression from cDNA arrays. Nevertheless, standard normalization resulting in ratios of the intensities was also performed in the separate experiment normalization scheme for comparative purposes.

The scheme in Figure 3.2 presents the typical course of a comparative analysis of expression values from two experiments. The diagram in Figure 3.3 illustrates the work flow for an initially proposed integrative approach strategy.

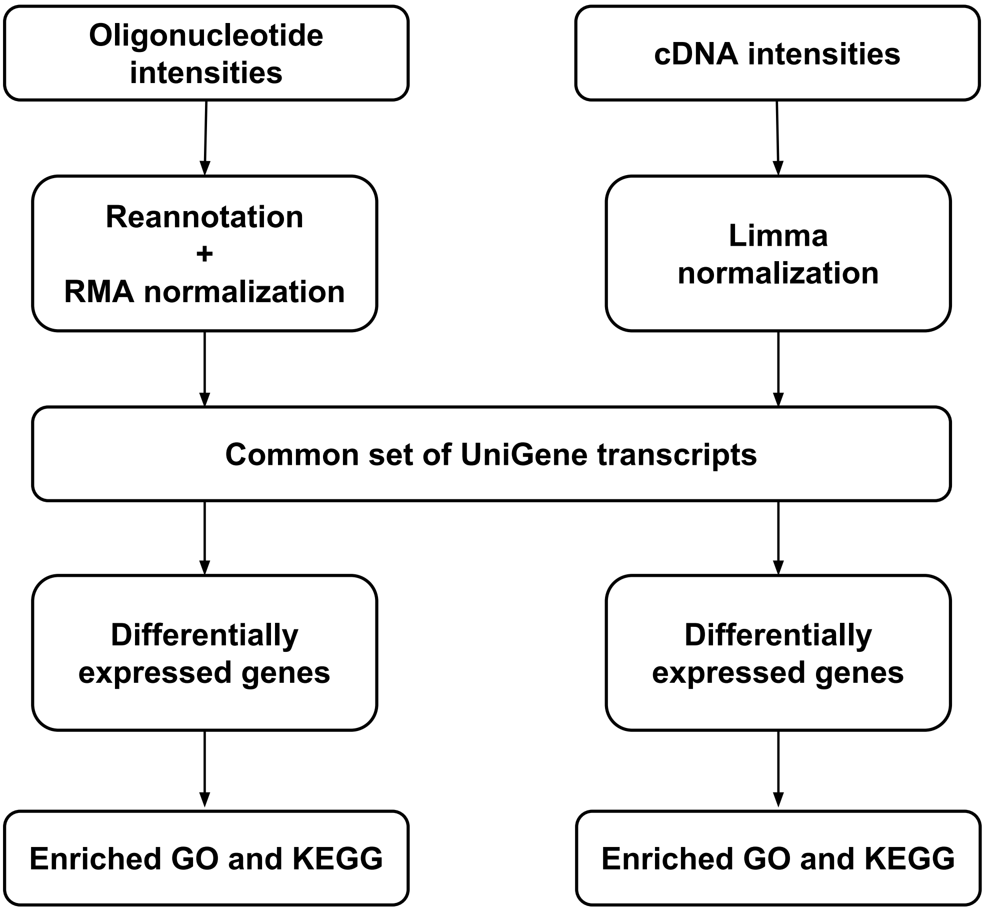

Oligonucleotide intensities
cDNA intensities
Reannotation + RMA normalization
Limma normalization
Common set of UniGene transcripts
Differentially expressed genes
Differentially expressed genes
Enriched GO and KEGG
Enriched GO and KEGG

Figure 3.2: Workflow for a standard microarray comparative analysis.

3.2.3 Different microarray platform batch effect correction

The goal of this work was primarily to carry out a combined analysis of the data. Therefore, the first step of data set integration was to extract a set of genes common for both platforms. This was accomplished on the basis of UniGene identifiers. Then on the common gene sets, batch effect correction through empirical Bayes methods was applied using ComBat (Johnson et al., 2007) software provided in the R SVA package for three batches (one for each of the two channels in cDNA data and one for oligonucleotide data) with no covariates. This lead to the transfer of expression values to a unified scale (Figure 3.4). Since the red and green channel data have been filtered for batch effects, their expression was merged as for technical replicates.

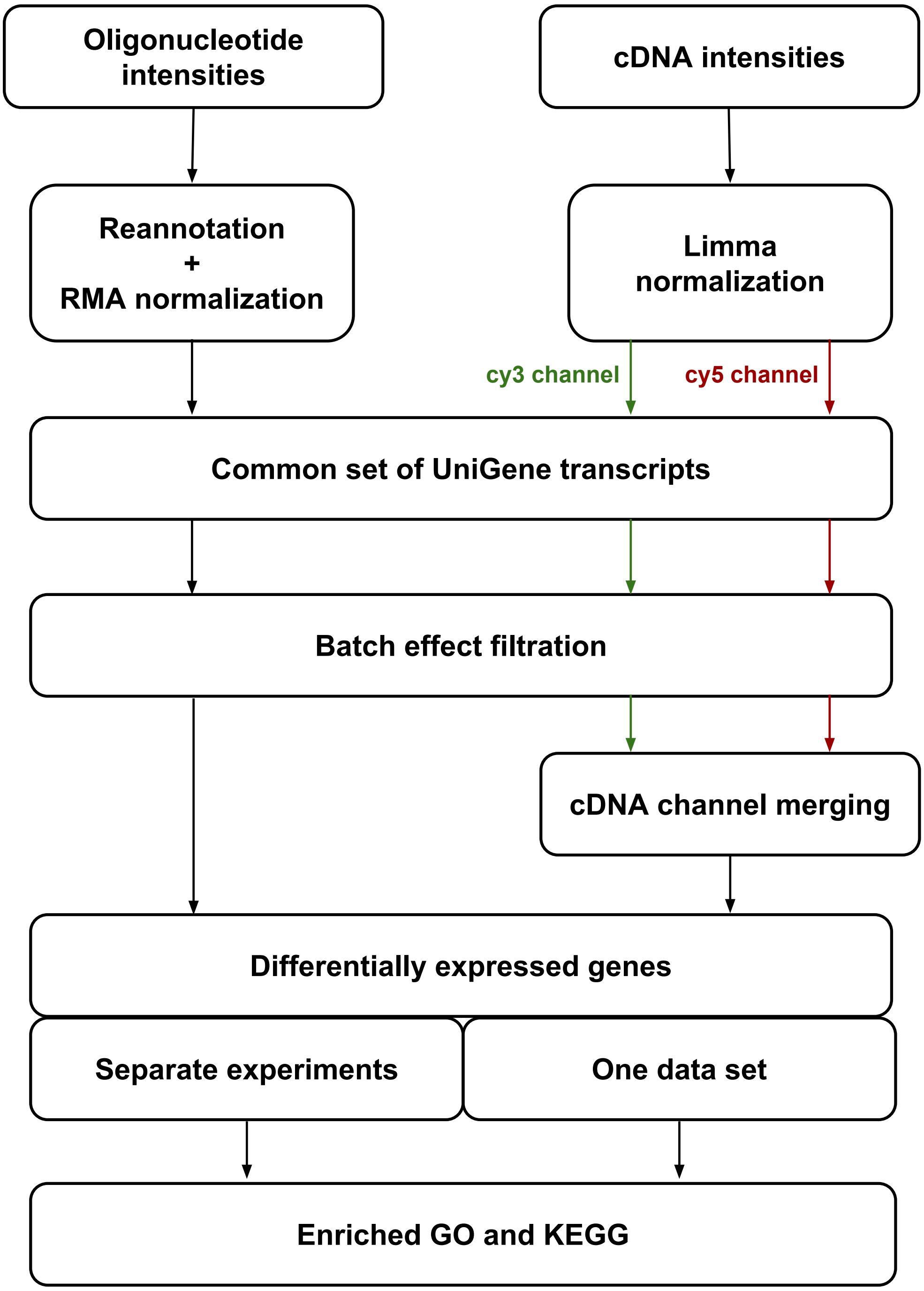


Figure 3.3: Diagram illustrating the proposed microarray initial data combination procedure.

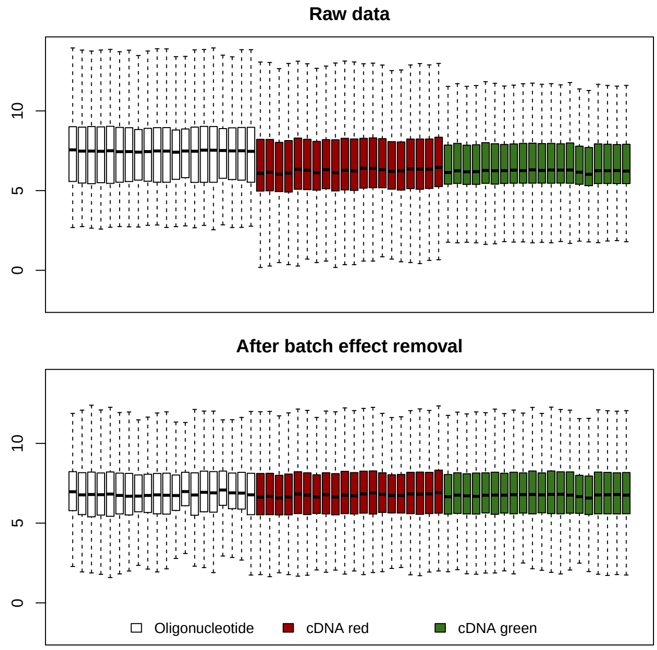


Figure 3.4: Exemplary sample distributions before and after batch effect correction.

3.2.4 Integrative approaches for radiosensitivity biomarker research

Statistical Analysis

As both in the cDNA and oligonucleotide experiments the absorbed doses may be classified as high according to (UNSCEAR, 2000), in the radiosensitivity biomarker analysis these samples were considered irradiated, regardless of dose. The genes in groups of irradiated and control samples were tested independently for differential gene expression with a statistical inference approach with the application of two-sample t-tests, modified Welch's tests or U-Mann-Whitney test, according to population normality and variance homogeneity assumption fulfillment. These tests were performed for both approaches: simple separate normalization of data from two experiments and the alternative unification of data using batch effect correction.

Moreover, after applying batch effect filtration the data sets could be considered numerically compatible, therefore, the samples for this method have been merged into one set and tested for differential expression.

The genes identified as differentially expressed were examined for ontology and signaling pathway enrichment in the GO (Ashburner et al., 2000) and KEGG (Kanehisa and Goto, 2000) databases.

Data combination approaches

In further investigation, as the data integration concept proved to be the correct course of action, three diverse techniques for data combination were adopted and compared for their performance (Papiez et al., 2015), denoted henceforth as:

- Restrictive
  The data sets were preprocessed and analyzed downstream independently, which resulted in lists of genes that were labeled as differentially expressed i.e. their p-value from statistical testing falls below the threshold of 0.05. Validation of the results from two single studies in this case assumes the form of the intersection of differentially expressed genes being considered the final gene list.
- Arraymining
  Data were analyzed independently, likewise to the restrictive approach. However, non-statistical data mining algorithms available on the Arraymining webservice (Glaab et al., 2009) were used. The gene signature was then chosen based on genes ranked as the most significantly differentiating in both experiments in the Ensemble of four methods: Empirical Bayes moderated t-test (Lönnstedt and Speed, 2002), Partial Least Squares cross-validation (Hall, 1999), Random Forest Mean Decrease in Accuracy (Breiman, 2001) and Significance Analysis of Microarrays (Tusher et al., 2001). The Ensemble forms a final gene list taking into account the sum of ranks for the individual algorithms.
- Integrative
  The method chosen here is based on weighted Z-scores p-value combination (Zaykin, 2011). The p-values from the two studies for each gene are joined after transformation with the inverse cumulative standard normal distribution function (Eq. 3.10). The integration of values derived from two-sided tests imposes the need to transform the p-values according to the recorded effect direction, as originally this method was designed for right-tailed testing (Eq. 3.12).

$$Z = \frac{\sum_{i=1}^{k} w_i Z_i}{\sqrt{\sum_{i=1}^{k} w_i^2}} \tag{3.10}$$

$$p = 1 - \Phi(Z) \tag{3.11}$$

$$p_{one\_sided} = \begin{cases} p_{two\_sided}/2, & \text{if effect direction} > 0 \\ 1 - p_{two\_sided}/2, & \text{otherwise} \end{cases} \tag{3.12}$$

The obtained Z-scores are combined with the weights set to the inverse standard error and transformed back to the form of a resulting p-value (Eq. 3.11). This procedure is presented in Figure 3.5. The eventual features with statistically significant combined p-values establish the final gene list.

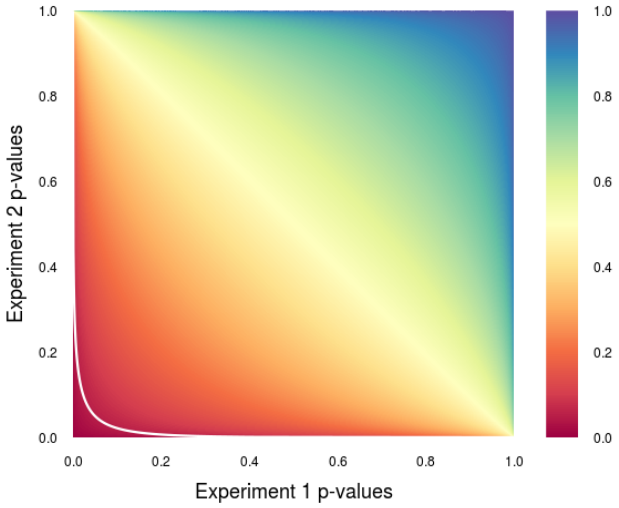


Figure 3.5: Illustration of the weighted Z-score p-value combination method. The values on the axes represent p-values potentially obtained in two experiments for matching transcripts. The weight in this case is the inverse standard error. The color depicts the combined p-value level. The white line illustrates the 0.05 threshold for the resulting combined p-value. The features with combined p-value below the white line are considered statistically significant.

Separability validation

The performance of three data combination approaches was assessed with the separability of the data sets based on the obtained gene lists. Separability in this case was assessed with a logistic regression model (Antoniadis et al., 2003). The applied model selection technique for regularization was carried out by means of the likelihood ratio test. Logistic regression models were chosen as an appropriate tool for signatures obtained in the course of statistical inference.

As the Ensemble of methods available in the Arraymining service is not statistically-based, but rather a data mining approach, a classic Support Vector Machine was applied for comparison to the logistic regression classifier. Model selection for the SVM was performed through minimizing the error rates. In order to provide a sufficient coverage of samples per variable, a maximum of 20 gene features was examined in each case. Initially, the regularization issue was controlled by means of the binomial test, yet this method, being very strict, produced a cutoff at one feature in all of the three studied cases. As this issue provokes a large loss of information, the minimal error approach remained the method of choice for feature selection.

The separability results were measured with Receiver Operating Characteristics, and specifically by means of the Area under the Curve metric. The ROC curves smoothing is performed with the binormal algorithm. Moreover, positive and negative predictive value (PPV & NPV) measures are calculated and compared. The class separability thresholds are tuned based on the ROC curves Youden's index (Youden, 1950).

3.2.5 Dose profile based preselection for classification

Another research aspect that has been examined in these data sets was the selection of appropriate inter-platform integration techniques to examine the gene expression response to different doses. The irradiated samples in both experiments received high doses, nevertheless, the question arose whether combining information from the two experiments while retaining the information about different dose levels (2 and 4 Gy) will have a significant impact on the results obtained. Thus, the joint data from two experiments was subjected to further classification analysis of dose profiles.

Differentiation analysis

The common gene set for the oligonucleotide and cDNA platforms was extracted for further processing. The first stage of statistical inference performed was a standard procedure, where differentiation tests between the dose groups were conducted: t-test, modified t-test and U-Mann-Whitney test, according to the normality and variance homogeneity assumptions.

Moreover, as an additional criterion for selection, only genes which did not produce significant differences between controls in the two experiments were selected to be used in the next stages. These genes were assessed additionally for differentiation between 2 Gy and 4 Gy with distinction between radiosensitive and radioresistant samples. The genes with differential expression between doses specific for radiosensitive and radioresistant patients were then investigated towards their functional characteristics verified using overrepresentation analysis of biological process Gene Ontology terms (Ashburner et al., 2000). Overrepresentation was measured by means of Fisher's exact test implemented in the topGO R package (Alexa and Rahnenfuhrer, 2010) with Benjamini-Hochberg correction for multiple testing.

Trend testing

Accounting for diverse doses in the two experiments, the genes were additionally inspected for the existence of trends using the Jonckheere- Terpstra test (Terpstra, 1952; Jonckheere, 1954). In this case, the hypotheses present as follows:

$$H_0 : \Theta_1 = \Theta_2 = \ldots = \Theta_k \tag{3.13}$$

$$H_A : \Theta_1 \leq \Theta_2 \leq \ldots \leq \Theta_k \tag{3.14}$$

where Θ_i is the $i - th$ sample median.

This renders the test an equivalent of the Kruskall-Wallis test, yet for samples that may be sorted. The genes recognized as significant at the level of 5% with strictly increasing and decreasing trends were further analyzed for Gene Ontology term enrichment. The genes denoted strictly increasing/decreasing fulfilled the condition, that they did not attain significance in the monotonic trend.

Moreover, for the purpose of this study extensive research was necessary into the nature of the trends. Therefore, proceeding in this direction, the analyzed genes not presenting significant differences between controls in the two experiments, were classified into one of the six types of response profiles (Figure 3.6):

- irradiation related up-regulated
- irradiation related down-regulated
- dosimetry applicable up-regulated
- dosimetry applicable down-regulated
- high dose activation up-regulated
- high dose activation down-regulated

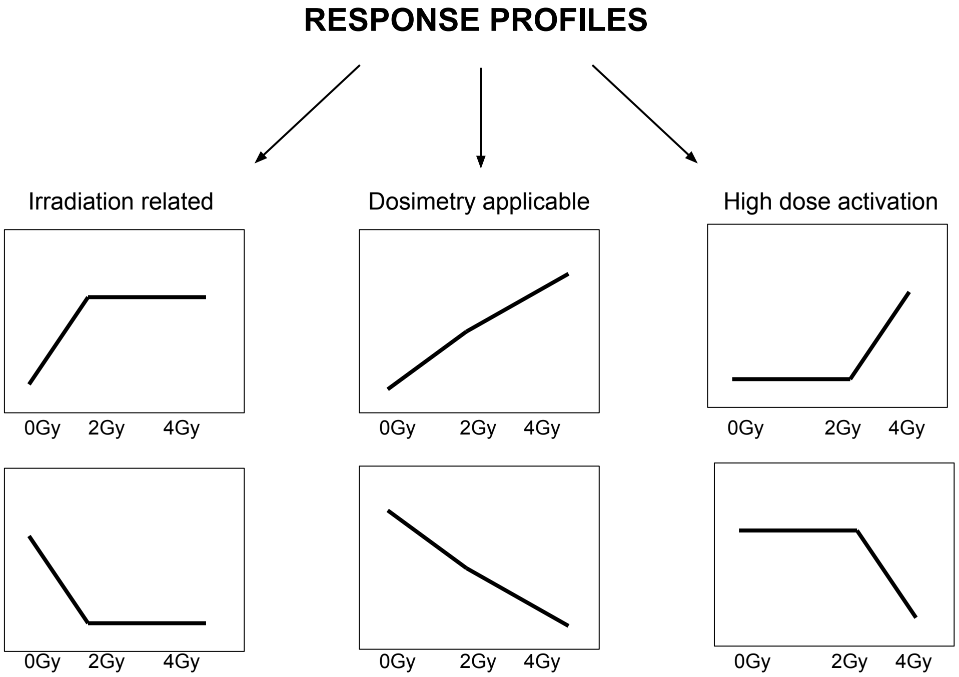


Figure 3.6: An illustration of dose response profiles applied for gene grouping to enable accurate expression value interpolation.

The labels refer to the potential utility of features falling into the group, i.e. the irradiation related genes are the ones activated by irradiating the samples on lower levels; the dosimetry applicable have a response profile changing with dose in the same direction and could be of potential use in dosimetry tasks; high dose activation profiles present a response only at the 4 Gy dose level, but none at 2 Gy. The above mentioned response profiles were designated based on the expression differentiation between doses, e.g. in the irradiation related up-regulated group a significant difference is observed between 0 Gy and 2 Gy, but there are no expression levels significantly differentiating between 2 Gy and 4 Gy.

Multiple random validation

Considering the identification of potential biomarkers of radiation response, the samples were classified by means of a multiple random validation procedure. Notwithstanding, due to the inconsistency between doses used in two experiments, simple separation between controls and irradiation samples was not attainable. Hence, information obtained by means of the trend testing stage was used and the ensuing procedure was performed on genes determined as belonging to the irradiation related and dosimetry applicable categories.

In the case of genes which were assigned to the dosimetry applicable group, expression values in the 2 Gy dose point were replaced with a linear interpolation value between the control and 4 Gy values in the corresponding samples. In the irradiation related group, as there were no significant differences between values in 2 Gy and 4 Gy dose points, the values remained the same. In this fashion, the desirable data set with two classes: controls and 2 Gy samples, was approximated. Results gathered through the course of validating this novel method were juxtaposed against multiple random validation performed on unadjusted expression values.

The multiple random validation scheme was executed in 500 repetitions. For each repetition the data were randomly assigned into training and test sets with a ratio of 7:3 and case/control proportions were retained at the level of the true ratio in the entire set. Logistic regression was chosen as a classification technique with forward stepwise feature selection using the Bayes Factor (Berger and Pericchi, 1996) as a criterion for increasing the number of model features. Genes forming the final model were recorded in each iteration, and later ranked according to the frequency of their occurrence in a single signature. The resulting list provided a reference for subsequent comparative analyses.

3.2.6 Monte Carlo Feature Selection validation

In order to validate the results of multiple random validation, the entire data set was subjected to distributed Monte Carlo based feature selection (MCFS), in order to identify genes showing the most significant interaction networks in terms of radiation response. This was achieved using the Broadside tool (Krol, 2015), which is a distributed feature selection and interaction mining algorithm and application designed for machine learning problems. The principle behind it is that interactions are captured by permuting pairs of variables, intercepting the effect these permutations have on the model performance measure, and solving linear equation systems to enable the performance of a decomposition of feature total effects into main effects and interaction effects. Respectively, Broadside is not bound to a specific type of model, which induces robustness, as it eliminates the risk of misinterpreting unpruned decision tree structures as valuable features and interactions. The most frequent genes in the multiple random validation logistic models were compared to the results of the Broadside MCFS networks.

3.3 Inter-omics data integration

The statistical p-value integration methods offered promising results in the attempt to perform combined analysis on different microarray platform datasets within the same transcriptomics space. This lead to the undertaking of a transfer of these procedures for the purpose of investigating two datasets from different fields: proteomics and transcriptomics on subgroups of workers from a nuclear production facility. The aim of this study was to further deepen the knowledge of molecular mechanisms related to radiation-induced human heart pathology.

3.3.1 Data sets

Proteomics samples

Left ventricle cardiac samples were extracted post mortem from 29 male individuals as described in detail in (Azimzadeh et al., 2017). These were exposed to different external doses of ionizing radiation during their lifetime. Non-exposed individuals of the same area were used as the control population. Both controls and exposed workers died of ischemic heart disease. The samples were categorized into four groups conforming to the total dose of ionizing radiation to which the individuals were exposed. The groups present themselves as follows:

- unexposed controls (3 samples),
- < 100 mGy low dose exposed (6 samples),
- $100 - 500$ mGy medium dose exposed (10 samples)
- > 500 mGy high dose exposed (10 samples).

The low number of individuals in each group was caused by difficulties in obtaining human data from the workers who died of the very specific ischemic heart disease condition and time of tissue collection (maxiumum 4h post mortem) was paramount in this case. Nevertheless, small sample sizes present a certain problem in the power of statistical procedures and require careful and accurate choice of analysis workflows.

The samples were processed in an LC-MS/MS experiment as reported previously (Azimzadeh et al., 2017). The clinical data available for the samples contained total external dose, age, smoking habits, alcohol consumption, and BMI. However, while the dose and age factors differed between the workers, all of the workers were recorded to be smokers and drinkers.

RNA-seq samples

RNA samples were collected initially from 8 male subjects, a subset of the group previously analyzed in the proteomics approach (Azimzadeh et al., 2017). They were divided into two groups: unexposed controls (3 samples) and $>$ 500 mGy high dose (5 samples). The mirVana PARIS Kit (Ambion, ThermoFisher, USA) was used to isolate both native protein and total RNA. Total RNA was isolated from the lysate according to the Ambion, ThermoFisher manufacturer's protocol. RNA integrity was assessed on the Agilent 2100 Bioanalyzer. The sequencing validation analysis supplied good quality data for 4 samples derived from the group of workers used in the proteomics experiment: 2 controls and 2 high-dose samples. The sequencing was executed on the Illumina NextSeq 500 desktop sequencer.

3.3.2 Confounding factor filtration

The first challenge posed by the data was the fact that there existed a strong correlation between age and dose factors in the studied individuals. As the total dose to which workers were exposed during their lifetime was the main point of interest in terms of the study of ischemic heart disease in this aspect, measures were introduced for the filtering of features that presented variability primarily due to the age factor. The dose-age correlation in the proteomics data was measured using Spearman's rank correlation coefficient. Furthermore, to avoid bias caused by the confounding age factor, a regression analysis with backward stepwise model building was performed. Each protein feature was modeled gradually excluding the factors of dose and age transformed using Box-Cox algorithm (Box and Cox, 1964) and model selection was conducted on the basis of Akaike Information Criterion (Akaike, 1974). Separability of the proteomics samples with the selected features was investigated using hierarchical clustering with Spearman's rank correlation as a similarity measure.

3.3.3 Statistical analysis of omics data sets

After the protein features explained mainly by the dose factor were extracted, primary analysis of the proteomics and transcriptomics data sets was carried out. Firstly, in the case of MS data, outlier detection was performed using Dixon's criterion within the dose groups. Then, within group normality was tested using Shapiro-Wilk procedure (Shapiro and Wilk, 1965) and based on the assumption notwithstanding, the Kruskal-Wallis test (Kruskal and Wallis, 1952) with Storey's FDR multiple testing correction (Storey, 2002) was used to assess differentiation among the dose groups. As a post-hoc method, Dunnett's test (Dunnett, 1955) was selected for determining deregulated proteins among the dose groups in relation to the control samples. Significance in the above tests was assumed at the level of 0.05 in all of the above mentioned procedures.

Transcriptomic RNA-seq data preprocessing was performed using state-of-the-art methods. Alignment and mapping were accomplished with STAR software version 2.5.1 (Dobin et al., 2013) against the GRCh38/hg38 human reference genome. Sorting and indexing was executed using SAMtools, version 1.3.1 (Li et al., 2009). The correlation between biological replicates within control and high-dose groups was assessed. Further differential expression analysis was performed with the use of R DESeq2 package (Love et al., 2014) with gene expression modeled based on the negative binomial distribution.

3.3.4 Functional analysis

Enrichment analysis of deregulated genes and proteins was carried out including Gene Ontology Biological Process terms and KEGG signaling pathways. Overrespresentation was assessed using Fisher's exact test. Moreover, gene and protein interaction and signaling networks were analyzed through the STRING search tool (http://string-db.org).

3.3.5 Statistical integration

Finally, after separate processing completion, an integrative multiomics analysis of data from the workers samples was conducted in the form of Fisher's combined p-value transformation(Fisher, 1992) on the common in both data sets gene and protein features (Figure 3.7). The combination is achieved by summing k log-transformed individual p-values. The inverse sum multiplied by 2 becomes the combined statistic, which follows the χ^2 distribution with $2k$ degrees of freedom (Eq. 3.15).

$$X = -2\sum_{i=1}^{k} log(p_i) \tag{3.15}$$

$$X \sim \chi^2_{2k} \tag{3.16}$$

This method is adequate for sequencing data, as the read counts cannot be approximated with a Gaussian distribution. Regarding the proteomics data, high-dose group samples and controls were only taken into account for combined analysis to ascertain compatibility with the transcriptomics data. Bearing in mind the non-specific nature of gene-protein coupling (multiple genes may correspond to a single protein), in such cases genes with the minimum p-value were considered. The combined p-values were afterwards corrected for multiple testing using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995).

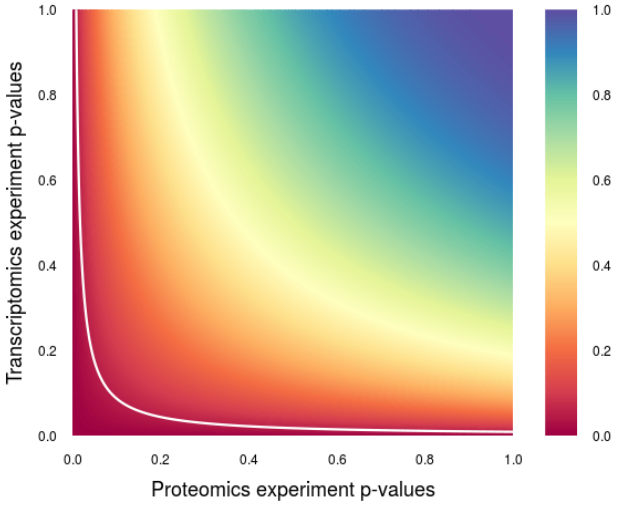


Figure 3.7: Illustration of the Fisher's p-value combination method. The values on the axes represent p-values potentially obtained in two experiments for matching features - in the case of this study proteins and transcripts. The color depicts the combined p-value level. The white line illustrates the 0.05 threshold for the resulting combined p-value. The features with combined p-value below the white line are considered statistically significant.

3.4 Different tissue and dose proteomics data integration

3.4.1 Data

Three biological replicate samples were collected from each of the four cell systems:

- Human skin fibroblasts
- Human coronary artery endothelial cells
- Human mammary epithelial cells (MCF10A)
- Human leukocyte cells

The cell exosomes were then divided into four dose groups, according to the radiation they were subjected to: 0 Gy controls, 1 Gy, 2 Gy, 6 Gy. Furthermore, the samples were processed using LC-MS/MS for protein identification.

3.4.2 Integrative tissue analysis

The exosome data were analyzed initially for protein identification. Afterwards, the common set of proteins from all tissues was examined in terms of clustering and similarity. Unsupervised hierarchical clustering was performed, followed by the determination of similarity indexes between individual samples (Frank et al., 2007):

$$SI = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 y_i^2}} \tag{3.17}$$

where x_i and y_i are abundance levels of proteins from two measured samples.

After examining the similarity between samples, deregulated protein analysis was performed. Differentiation of proteins between doses with regard to the control samples was determined by means of Dunnett's test at a 5% significance level within individual tissue groups. Moreover, the differentiating proteins were compared between the tissues.

Chapter 4

Results and Discussion

4.1 Batch effect identification

Firstly, in this section batch structure identification results accomplished using the dynamic programming algorithm are presented. Moreover, correction based on combining the novel algorithm of batch effect identification with an algorithm for batch bias removal is discussed. The choice of batch effect correction algorithm was carried out based on comparative studies (Chen et al., 2011; Luo et al., 2010), which conclude that the ComBat algorithm (Johnson et al., 2007) is a reliable, state-of-the-art method for batch effect removal, presenting in most cases the best quality of results compared to other approaches. Therefore, the ComBat method is used as a tool for correction, combined with the BatchI method for batch effect identification.

The experimental data with known *a priori* batch structure: E-GEOD-19419 and E-GEOD-36398, RNA-seq data and MS data, were first analyzed in terms of estimating the accuracy of the known, true structure of batches obtained when applying the dynamic programming BatchI algorithm (Papiez et al., 2018b). Furthermore, in these data, the quality of batch effect filtration was assessed in terms of intragroup correlation. In this sense, it is assumed that when batch effect is correctly identified and corrected for, the technical sources of bias are filtered from the data, which is expected to increase the correlation between samples within a studied biological condition group.

Finally, microarray expression sets E-GEOD-2034, E-GEOD-4183 and E-GEOD-10927, with unknown batch status labeling were processed. In each of the analyzed sets, the date of the experiment was known and used as the sorting factor. In order to evaluate the obtained results after batch effect identification and correction, the intragroup correlation index was used and the increase of Information Content of gene ontology terms enriched with differentially expressed genes was measured.

4.1.1 Known structure of batches

Batch division re-identification

The division into batches using BatchI dynamical programming was compared with the original batch grouping with the use of weighted average pairwise Dice-Sorensen Index (Dice, 1945) for the purpose of measuring the efficacy of batch effect identification. The Dice Index reflects the similarity of two data sets, with a value of 0 when there are no common elements, to a value of 1, when the sets are identical. Comparisons of true and identified batch structures are illustrated in Figure 4.1. True batches are presented by means of different symbols and colors, while the estimated structure is depicted by vertical lines dividing samples into batches.

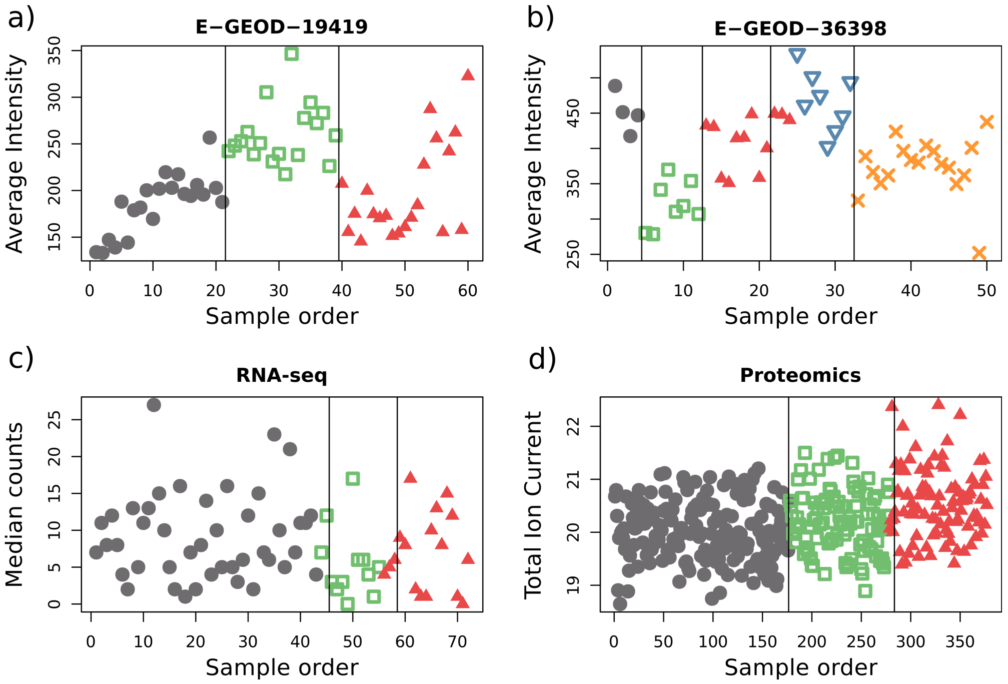


Figure 4.1: Division of the data sets into batches with the a priori defined groups and determined with the dynamical programming approach: a) Set E-GEOD-19419, b) Set E-GEOD-36398, c) RNA-seq data, d) Proteomics data. Colors show the original batch structure, the vertical lines present divisions found using the dynamic programming algorithm.

- Microarray data

  In the E-GEOD-19419 experiment the reproduction of batches is identical to the original division. In E-GEOD-36398 the batch assignment reconstruction is also highly accurate with a weighted average Dice Index of 94.05%. The fault is only in three samples belonging from third batch being assigned into the fourth.

- RNA-seq data

  In the RNA sequencing data the original batches are reconstructed with the value of a weighted average Dice Index of 93.02%. Two samples from batch no.2 were assigned to batch no.1 and three samples from batch no.3 to batch no.2.

- Mass spectrometry data

  In the case of MS data, batches are mapped with a weighted average Dice Index value of 99.78%. One of the samples from batch no.1 and five from batch no.3 were classified as batch no.2.

Batch effect correction

The data sets examined for the purpose of algorithm performance assessment when batch effect was previously identified and on record were evaluated in two aspects.

First, intragroup correlation was measured for samples, which belong to one biological condition investigated in the study. 95% confidence intervals for mean Spearman's correlation coefficients were computed and are depicted in Figure 4.2. For the two gene expression microarray experiments a significant increase in intragroup correlation after batch effect removal is evident. The RNA-seq experiment, having a strong design imbalance when considering the number of samples in the particular biological conditions, was expected to not present weaker batch effect identification performance. Nevertheless, even in the less numerous groups mean correlation within groups does not decline significantly. In the mass spectrometry data, which in contrast to the previous experiments was obtained through MALDI-ToF measurements, which are a quantitative technique, there is a clear increase in within group correlation, though larger differences may be observed compared to the original batch structure correction.

Moreover, the δ gPCA statistic was utilized as another qualitative measure of

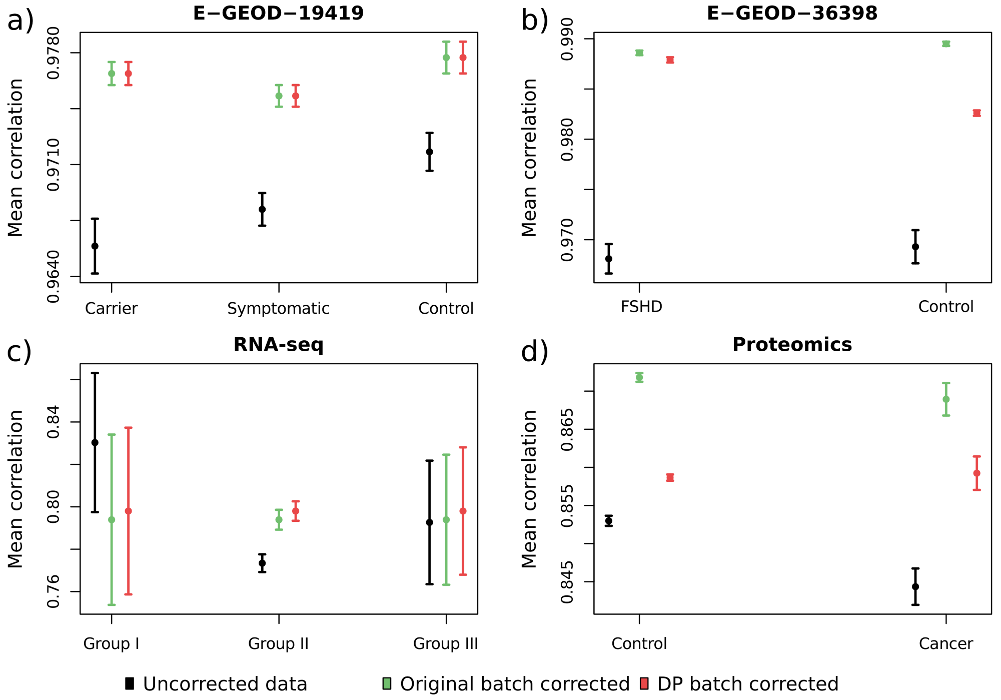


Figure 4.2: 95% confidence intervals for mean intragroup correlation coefficients in known batch structure data sets: a) Set E-GEOD-19419 (60 samples), b) Set E-GEOD-36398 (50 samples), c) RNA-seq data (72 samples), d) Proteomics data (373 samples).

change between data with mitigated batch effects versus no correction. The significance of this statistic with relation to no batch correction was evaluated using p-values estimated in the course of permutation tests (Table 4.1). In the gene expression microarray data, the change of δ gPCA statistic is significant in both experiments when applying batch effect correction based on the batch structure identified using the dynamic programming algorithm. In the RNA-seq data set the change after correction becomes significant when considering the structure information derived by means of the BatchI algorithm. In the MS data, with it being a large data set in terms of sample numbers, which contributes to the overall weak variation observed, there is no substantial difference after batch effect correction neither in the case of original batch labeling, nor the one derived using dynamic programming.

4.1.2 Detecting and correcting batch effect of unknown structure

The three experiments selected for the assessment of batch effect identification without prior knowledge of batch structure were analyzed in the same manner as the studies with *a priori* known batches of samples in terms of partitioning quality. This consists of

Table 4.1: Percent of variation induced by batch effect with regard to total variation, the corresponding gPCA δ statistics and the p-values for testing the significance against no batch effect correction for two microarray, an RNA-seq and a proteomics data sets.

E-GEOD-19419	**Original batch corrected**	**DP batch corrected**
Total variation [%]	69.23	69.23
δ	0.9271	0.9271
p-value	4.69E-08	4.78E-08
E-GEOD-36398	**Original batch corrected**	**DP batch corrected**
Total variation [%]	48.15	50.14
δ	0.9991	0.9989
p-value	2.24E-07	2.90E-07
RNA-seq	**Original batch corrected**	**DP batch corrected**
Total variation [%]	65.12	67.23
δ	0.2765	0.6175
p-value	4.87E-01	9.38E-02
Proteomics	**Original batch corrected**	**DP batch corrected**
Total variation [%]	23.82	24.56
δ	0.6645	0.6671
p-value	7.32E-01	7.15E-01

examining mean correlation within case/control subgroups. The results presented in Figure 4.3 indicate the fact that data integrity within analyzed biological groups is improved by including batch effect identification and subsequent correction steps. Likewise as in Figure 4.2, the errorbar plots are constructed on the basis of mean correlations and 95% confidence intervals. In all the three data sets E-GEOD-2034, E-GEOD-4183 and E-GEOD-10927, the use of batch effect correction executed with the dynamic programming algorithm and ComBat algorithm leads to the increase of the intragroup mean correlation for every one of the total of nine biological groups analyzed. In three of the nine groups of samples, a highly statistically significant increase is observed. Moreover, the proportion of variance explained by batch effects is decreased, which is reflected within the values of the δ gPCA statistic (Table 4.2). In the breast cancer experiment (E-GEOD-2034), six batches have been identified as the optimal number by the dynamic programming algorithm. For the colon cancer experiment (E-GEOD-4183): two batches, and for adrenocortical carcinoma (E-GEOD-10927): three batches.

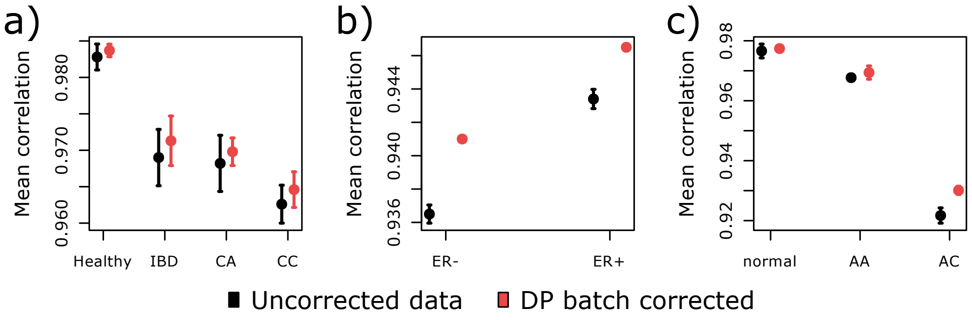


Figure 4.3: 95% confidence intervals for mean intragroup correlation coefficients in unknown batch structure data sets: a) Set E-GEOD-4183 (53 samples), b) Set E-GEOD-2034 (286 samples), c) Set E-GEOD-10927 (65 samples).

Table 4.2: Values of the gPCA δ statistic for different numbers of batches in the unlabeled data sets. The optimal number of batches is chosen with the minimum p-value principle (numbers in bold).

Breast cancer							
	2 batches	3 batches	4 batches	5 batches	**6 batches**	7 batches	8 batches
Tot. Var [%]	88.63	86.69	85.31	85.83	**84.10**	81.09	81.90
δ	0.45	0.43	0.44	0.53	**0.56**	0.56	0.51
p-value	6.89E-02	9.95E-02	1.05E-01	7.18E-02	**6.59E-02**	8.33E-02	1.48E-01
Colon cancer							
	2 batches	3 batches	4 batches	5 batches	6 batches	7 batches	8 batches
Tot. Var [%]	**80.24**	64.51	61.63	55.13	54.42	50.76	37.30
δ	**0.49**	0.39	0.56	0.58	0.64	0.68	0.60
p-value	**2.42E-01**	6.27E-01	3.63E-01	4.37E-01	3.97E-01	3.71E-01	7.02E-01
Adrenocortical carcinoma							
	2 batches	**3 batches**	4 batches	5 batches	6 batches	7 batches	8 batches
Tot. Var [%]	82.48	**79.04**	74.09	69.26	66.41	54.73	37.13
δ	0.45	**0.56**	0.57	0.56	0.54	0.51	0.62
p-value	1.46E-01	**7.61E-02**	1.60E-01	2.47E-01	2.28E-01	4.09E-01	3.40E-01

Runtime analysis

The E-GEOD-2034 data set, being one of the more numerous (286 samples in total), was selected additionally for the purpose of measuring algorithm runtimes. Firstly, the dynamic programming algorithm was run to scan for the optimal number of batches between 2 and 10 on the entire data set, and next on subsets comprising consecutively 80%, 60%, 40% and 20% of the samples. This approach was re-iterated 5 times in order to measure dispersion of the accomplished runtimes. Testing was performed in parallel mode on Intel®Core^{TM}i5-3320M CPU @ 2.60GHz × 4 processors. Results, presented in Figure 4.4, show that the method requires linearly increasing times with increased sample size, with an overall runtime reasonably small, even on a personal computer.

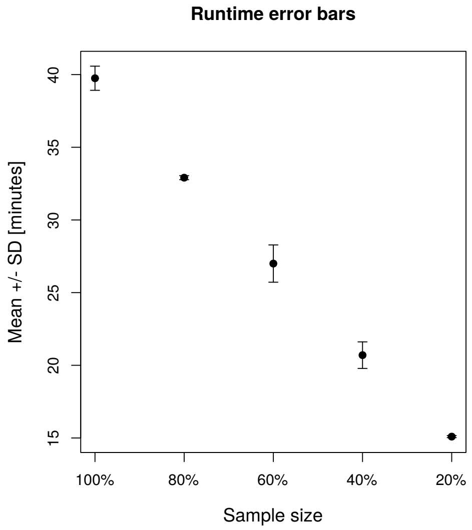


Figure 4.4: Runtime error bars showing the linear dependency between data set size and runtime. The original data set consisted of 286 gene expression microarray samples.

Functional gene ontology analysis

The data sets with unknown *a priori* batch structure were then examined in terms of functional analysis using GO terms in order to determine the relevance of biological conclusions, which may arise from the experiments. The differentially expressed genes identified in the original data sets with and after batch effect correction were used for GO term enrichment analysis by means of the hypergeometric test. The resulting lists of terms were afterwards compared and terms unique to each analysis workflow were thoroughly investigated.

In the E-GEOD-10927 experiment, which is a study on adrenocortical carcinoma and adenoma the enriched terms were matched with literature knowledge on these processes. The findings elucidated the irrelevance of GO terms unique to the lack of batch effect correction approach to the studied medical case. However, a majority of the GO terms gained by means of including batch effect filtration in the preprocessing has previously proven links to processes related with adrenocortical carcinoma and adenoma (Full list in (Papiez et al., 2018b)).

The remaining two studies concerned more well-defined and studied biomedical problems, i.e. breast and colon cancer, and therefore, the resulting GO term lists were large. In this case instead of literature studies the biological value of the findings is presented by means of the Information Content (IC) measure (Resnik, 1995). The assumption behind using this method is that, when batch effect correction is performed, a more detailed representation of the studied process is obtained, which is equivalent to an increase of IC value (Figure 4.5).

Additionally, when examining the dynamic programming combined with ComBat correction data sets, functional analysis results have been compared with GO terms acquired with data corrected using an alternative SVA approach (Leek and Storey, 2007). This method consists of identifying existing batch effect variability in the data and simultaneous filtration of the effects based on the estimated model.

Preserving the different branch and node sizes in the Gene Ontology graph, the total IC measure was standardized per GO term. The results prove that when it comes to common well described diseases, such as breast cancer (incidence rate $200 - 900$ cases per million (Ferlay et al., 2015)) or colon cancer (incidence rate $50 - 400$ per million (Haggar et al., 2009)), preprocessing data with the dynamic programming approach does not lead to a significant improvement in the quality of the information (reflected by standardized total IC). However, this confirms that though preprocessing methods, including batch effect identification and correction, are essential for careful data analyses, they alone are not sufficient to provide an augmentation of the biological knowledge available in bioinformatics data bases for well described diseases. Still, when examining the less prevalent case of adrenocortical carcinoma ($0.5 - 2.0$ cases per million (Kerkhofs et al., 2013)) data processed using the BatchI dynamic programming algorithm provides a supreme outcome, which elevates the chance of discovering potential new mechanisms of disease. Furthermore, in each of the presented cases the BatchI identification approach combined with ComBat correction gives higher standardized total IC values than the alternative SVA method and the outcome is no worse when confronted with the uncorrected data (Figure 4.5).

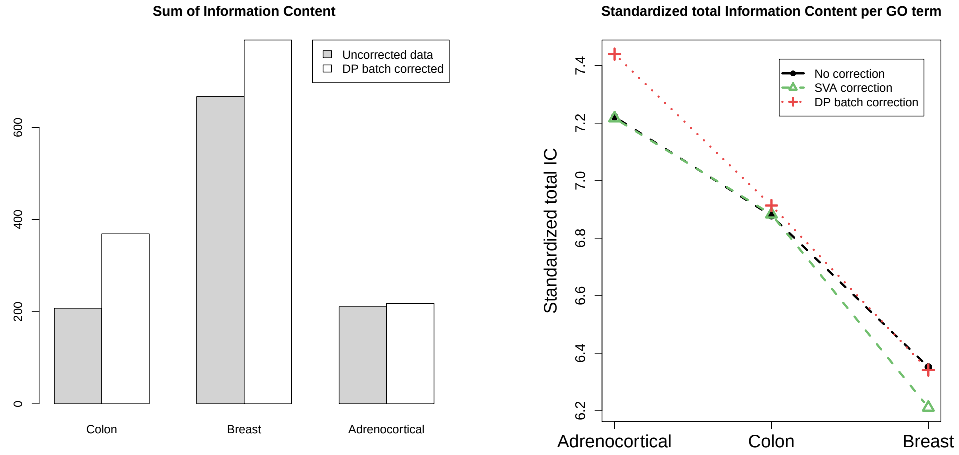


Figure 4.5: Comparison of Information Content for three studies. On the left Total Information Content of ontologies is presented for genes unique for data without batch processing and including batch effect identification with dynamic programming and correction. On the right standardized Information Content per GO term of ontologies for genes unique for data without batch processing and including batch effect identification and correction.

4.2 Inter-platform transcriptomics data integration

The data from two microarray experiments conducted on different platforms were combined with the aim of preserving coherence in the biological sense. In this way raw intensity signals for the cDNA data were extracted for each of the two color channels: red and green dyes. Afterwards, batch effect correction was applied and for every feature in the individual samples the intensity was averaged over the two dye-swap replicates. Next, the intersection of transcripts common for both biochips based on UniGene identifiers was subsequently analyzed. The numbers of genes retained for further research is presented in Figure 4.6.

4.2.1 Identification of Differentially Expressed Genes

Statistical testing was carried out initially for genes from samples, which were normalized separately in the two experiments, and samples processed with batch effect correction. These tests were performed with regard to the radiosensitivity status of the studied patients in order to identify differentially expressed genes (RR vs. RS). Then the separate analysis was compared to the results obtained by means of processing data from the two experiments combined as if it were one sample. The numbers of differentially expressed genes for control samples is presented in Table 4.3, and for irradiated

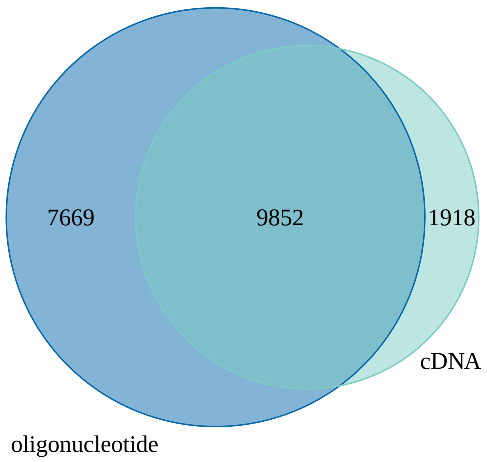


Figure 4.6: Venn diagram illustrating the proportion of genes common for both microarray platforms.

in Table 4.4.

	(A) **separate** **normalization**	**(B)** **batch effect** **adjustment**	**(A∩B)** **intersection**
oligonucleotide	577	577	577
cDNA	922	1093	380
Common	44	53	12
One data set	–	3146	–

Table 4.3: Number of differentially expressed genes at the significance level of 5% for control samples.

The results show that when studying the data from the two microarray experiments, usually the procedure comprising batch effect correction yields a larger number of DEG. Moreover, the situation where the expression sets are merged into one provides notably considerable numbers of genes classified as differentially expressed, yet this is due to a lack of specificity and a raised probability of retrieving false discoveries when increasing sample sizes. This shows that straightforward merging of datasets is not an adequate tool for integrative data analysis.

However, as the number of DEG is not a sole measure of interest in biological studies, but rather the relevance of the results towards elucidating studied processes, the DEG common for the two studies, which were obtained as a result of separate data

	(A) **separate** **normalization**	**(B)** **batch effect** **adjustment**	**(A∩B)** **intersection**
oligonucleotide	633	633	633
cDNA	669	1159	289
Common	38	51	12
One data set	–	3526	–

Table 4.4: Number of differentially expressed genes at the significance level of 5% for irradiated samples.

analysis, and merging the data into one set, were tested for statistically significantly enriched ontologies and pathways in the GO and KEGG databases, using the hypergeometric test. Ontologies and pathways were identified as significantly enriched at the significance level of 5%. Essentially, the differentially expressed genes obtained in the course of the batch effect correction approach were linked to a wider range of ontologies and pathways. Specific interest is drawn towards radiation induced related processes that have not been determined in the case of data normalized separately. For example, a strong group of processes and functions were linked with the MAPK signaling pathway which has been described as playing a key role in the molecular background of radiosensitivity (Chung et al., 2009). Moreover, annotations to the radiation-related p53 regulation (Mirzayans et al., 2013) and mTor (Steelman et al., 2011) pathways manifested themselves. Other annotations to ontologies including cellular response to stress, apoptosis and regulation of cell death may further point to a key role of the identified DEG and radiosensitivity.

4.2.2 Data integration approaches

In order to further investigate the approaches for data integration, results for three methods were compared in terms of differential gene expression and their utility towards separability between the radiosensitive and radioresistant patient groups. In this section the analyzed data is calculated as the signal log ratio between irradiated and control samples.

- Restrictive approach

  The two experiments produce lists of genes containing 471 and 927 DEG at the significance level of 0.05, for the oligonucleotide and cDNA experiments respectively. The intersection of these two sets consisted of 30 genes (Figure 4.7).

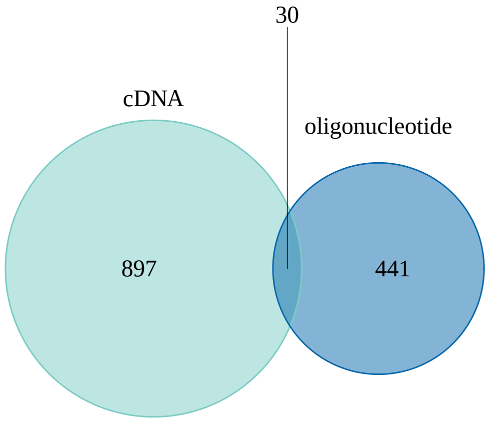


Figure 4.7: Venn diagram for differentially expressed genes in two experiments.

- Arraymining

  The procedures implemented in the Gene Selection section of the Arraymining platform provide a list of a predefined number of top-ranked genes. Thus, an intersection of a 1000 top genes in the two experiments was analyzed and the gene set sizes presented in Figure 4.8.

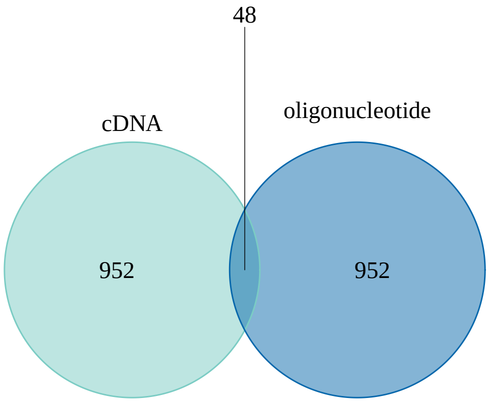


Figure 4.8: Venn diagram for Arraymining top ranked genes in two experiments.

- Integrative approach

  The p-value integration approach gene list was obtained by combining p-values using the weighted Z-score method. Weights were designated as the inverse standard error for the gene expression distribution. The approach resulted in a list of 108 differentially expressed genes significant at 0.05 level.

The sizes of common and unique lists, acquired in the course of comparing the three approaches, are illustrated in Figure 4.9. In total, 12 genes were common for all three data integration approaches.

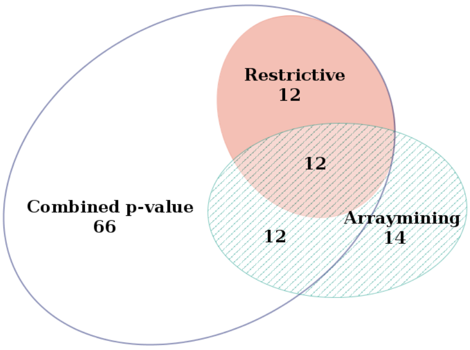


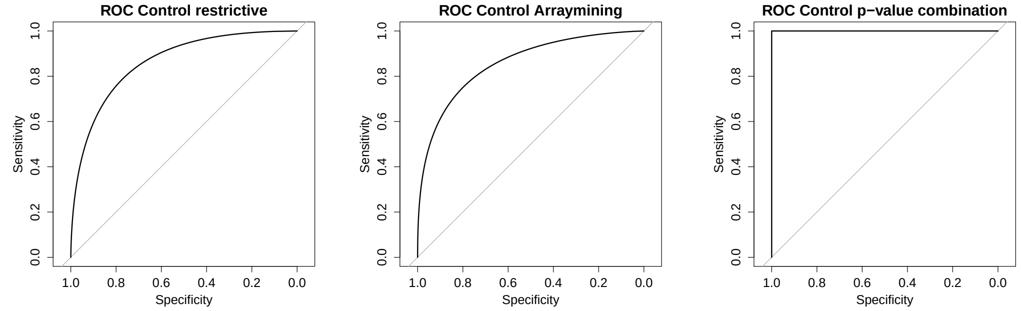
Figure 4.9: Venn diagram for gene lists obtained in three data integration approaches (Micallef and Rodgers, 2014).

This shows that integrating expression data at the p-value level is an adequate method for enhancing results in the form of a differentiating gene signature. Setting a fixed p-value threshold on the individual data sets often leads to a binary decision whereas taking into account the p-values tied to a test statistic enables a more precise incorporation of the differentiating strength of a particular gene feature and avoid the potential rejection of genes of interest, which would take place when setting an arbitrarily significance level.

4.2.3 Separability analysis

Logistic regression model

The DEG lists were used to construct a logistic regression model for measuring separability of the two groups of samples: radioresistant and radiosensitive patients. In each of the three data integration approaches model selection was carried out using the likelihood ratio test. The resulting models comprised 6, 6 and 16 features respectively for the restrictive, Arraymining, and integrative approaches. When applying the signature to discriminate between RR and RS patients, solely the integrative approach signature produces a model with perfect separability. The model efficiency comparison is presented with the use of Receiver Operating Characteristics (ROC) in Figure 4.10. Additionally, the decrease of error rates depending on the number of features is illustrated in Figure 4.11.

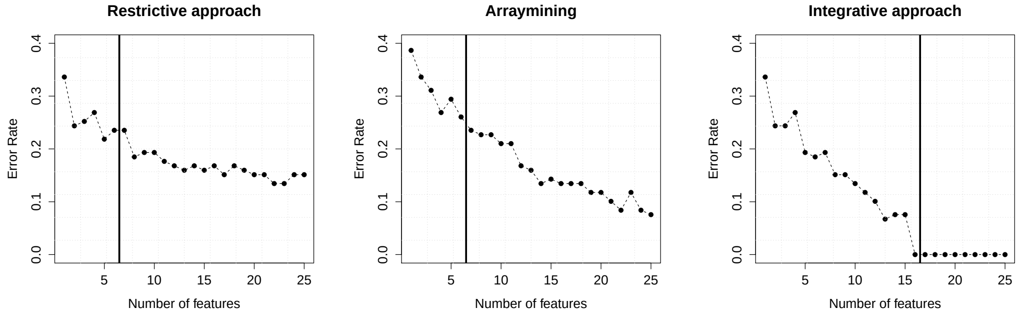


(a) Restrictive
AUC = 86.2%

(b) Arraymining
AUC = 85.6%

(c) Integrative
AUC = 100.0%

Figure 4.10: Receiver Operating Characteristic curves for logistic regression model separability.



(a) Restrictive

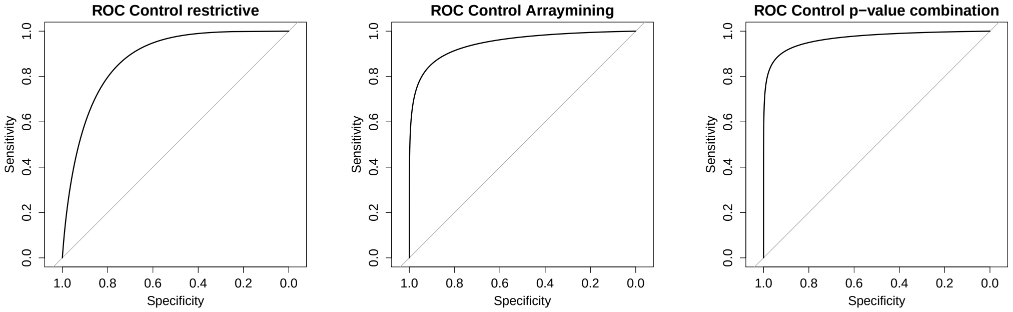
(b) Arraymining

(c) Integrative

Figure 4.11: Separability error rates subject to the number of features in the logistic regression model. The vertical line demonstrates the borderline, beyond which further feature addition does not provide a significant increase in model performance based on the likelihood ratio test.

Support Vector Machine

To challenge the assumption that a logistic regression model will perform better on features selected by means of statistical tools, another classification tool from the data mining field was chosen for comparison purposes: Support Vector Machine classifiers. It was tested on all three gene signatures. Model selection in this case was based on the minimum error rate for a given number of features. The three signatures provided in models built of 8, 19 and 19 features, respectively for the restrictive, Arraymining, and integrative approaches. The Receiver Operating Characteristics (Figure 4.12) suggest once more the superiority in terms of performance obtained by means of the p-value integration signature. Moreover, the model provides the lowest error rates (Figure 4.13).

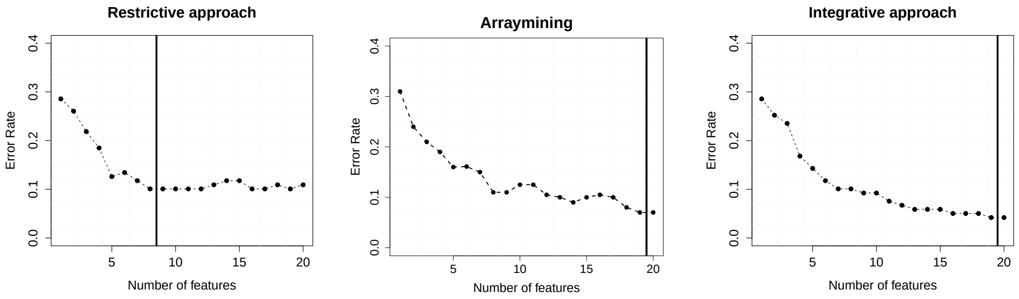


(a) Restrictive
AUC = 87.9 %

(b) Arraymining
AUC = 94.6 %

(c) Integrative
AUC = 96.7 %

Figure 4.12: Receiver Operating Characteristic curves for SVM model separability.



(a) Restrictive

(b) Arraymining

(c) Integrative

Figure 4.13: Separability error rates subject to the number of features in the support vector machine model. The vertical line demonstrates the borderline separating the optimal number of features minimizing the error rate.

The model performance focused on radiosensitivity group separability was also measured by comparing Positive and Negative Predictive Values. These statistics are summarized in Table 4.5.

	Logistic regression		Support vector machine	
	PPV [%]	NPV [%]	PPV [%]	NPV [%]
Restrictive	86.67	74.32	88.33	91.52
Arraymining	70.13	90.47	92.98	91.94
Integrative	100.00	100.00	98.18	93.75

Table 4.5: Positive and negative predictive value for logistic regression model and support vector machine separability.

The augmentation of gene signatures is not only a quantitative increase, but more importantly enhances the quality in terms of classification potential. The separability of the integrative approach model proved to be the single case of signature providing perfect distinguishing of the data groups coupled with the statistically based logistic regression model with feature selection performed by means of the likelihood ratio test. By contrast, the best model for the support vector machine is also obtained with the integrative approach signature, yet does not result in perfect separability. Moreover, when considering the error rate decrease, the integrative approach yields the lowest error rate values. Finally, the PPV & NPV prevail for both the logistic regression and the Support Vector Machine models.

Finally, the obtained signatures were validated in functional analysis by extracting the enriched KEGG signaling pathways. It revealed pathway terms for the integrative approach gene set that were not determined through the restrictive or Arraymining approaches. These include pathways connected with radiation exposure and susceptibility or cancer, such as JAK-STAT (Ding et al., 2013), T cell receptor (Witek et al., 2014), cytokine-cytokine receptor interaction (Herok et al., 2010), Fc epsilon RI (Fox et al., 1976) and natural killer cell mediated cytotoxicity (Son et al., 2014).

4.2.4 Dose response trend analysis

Beforehand, the focus of this study was to test for the ability to discriminate between radiosensitive and radioresistant patients. For this purpose, the two dose groups from the different experiments (2 Gy and 4 Gy) were analyzed together and assigned a high-dose label. However, the second research question that was to be addressed in this data set, is the radiation dose response patterns. In this case the analysis approach changes and the two different doses are considered separately, and the combination focuses on features where there is no significant difference between control samples in the two microarray experiments.

Differentiation analysis

Upon extracting genes, which are common for the two microarray platforms: Affymetrix oligonucleotide and custom cDNA, a unified procedure was carried out on a total of 9852 genes. Initially, the common genes were assessed concerning the control samples in order to provide the same base level for datasets from both platforms. Among the control samples in the two studies 7429 genes were not identified as significantly differentiating between the preprocessed data sets. Henceforth, this gene set was examined in order to identify features that display different patterns of response, separately for radiosensitive and radioresistant patients. The intersection of DEG significantly differentiating expression levels in the 2 Gy and 4 Gy dose groups is presented in Fig. 4.14. A total of 1214 genes was identified uniquely to the RS group and 730 genes to the RR group.

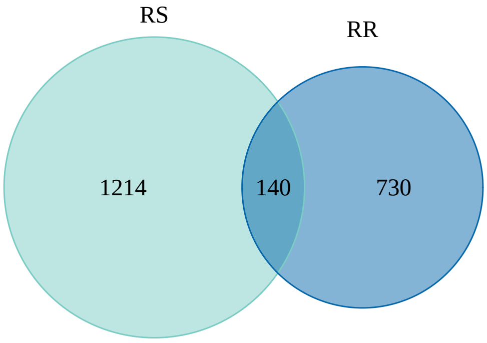


Figure 4.14: Venn diagram presenting a comparison of the numbers of genes differentially expressed between 2 and 4 Gy doses in RR vs. RS samples.

Next, Gene Ontology enrichment was analyzed by means of Fisher's exact test with regard to all of the differentially expressed genes between 2 Gy and 4 Gy. Benjamini-Hochberg multiple testing correction was applied and one significantly enriched GO term remained, i.e.: **cellular amino acid metabolic process**. For the radiosensitive group no significantly overrepresented terms were determined, yet in the radioresistant patients, 31 terms were statistically significantly overrepresented. The GO biological process terms enriched in the RR group included, among others radiation-induced mechanisms such as: stress response, oxidative phosphorylation, and immune response regulation.

This primary analysis of DEG subject to different doses of radiation suggested potential changes in the RR and RS breast cancer patients expression profiles. The DEG unique for the radioresistant group are involved in a variety of biological processes. Examples have been demonstrated, as previously reported, to play major roles in radiation response and tumor development (Weichselbaum et al., 1994; Park et al., 2014; Reinhardt et al., 1997). By contrast, the radiosensitive group gene are not overrepresented in biological processes known to be of biological importance. The findings are coherent with reports of key processes being silenced in radiosensitive patients and suggest an area of further experimental investigation.

Trend testing

The numbers of genes with increasing, decreasing and monotonic trends according to the results of Jonckheere-Terpstra test are presented in Table 4.6. Furthermore, genes were grouped as strictly increasing and decreasing if they did not show significance in the monotonic trend. Afterwards, the strictly increasing and decreasing dose response genes were analyzed for GO term enrichment. The strictly increasing genes were represented in 99 significantly enriched terms, and the down-trending in 38 GO terms. Some of the terms associated with decreasing genes feature processes related to hemopoiesis and homeostasis, GPI anchor metabolism and biosynthesis. By contrast, terms overrepresented by increasing trend genes were linked among others to cellular response to ionizing radiation and Wnt signaling. The latter has been reported to be linked to breast cancer mechanisms in a study comprising a large dataset analyzed in a non-standard manner investigating beyond differential expression (Schmid et al., 2012). The decreasing trend genes processes include hemopoiesis and homeostasis, which have been previously shown to play a role in stem cell injury from ionizing radiation (Shao et al., 2014). Moreover, GPI anchors being important apoptosis regulators when deregulated by ionizing radiation have a potentially significant impact on cellular resistance (Brodsky et al., 1997).

Table 4.6: Numbers of genes showing significant dose trend. The strictly increasing and decreasing genes are those which do not appear in the monotonic trend group.

	Increasing	**Monotonic**	**Decreasing**
N^o of genes	717	377	53
	Strictly increasing	**Strictly decreasing**	
N^o of genes	363	30	

For the classification task, selection of interpolation method between doses was proposed based on the gene profiles (Figure 3.6). The numbers of genes classified into six types of response profiles are summarized in Table 4.7.

Table 4.7: Numbers of genes grouped in to particular dose response profiles.

Number of genes in response profiles					
Irradiation related		Dosimetry applicable		High dose activation	
Up-No change	610	Up-Up	117	No change-Up	48
Down-No change	1067	Down-Down	969	No change-Down	319

Multiple random validation

In order to summarize the results of the multiple random validation procedure, average statistics were computed over the total of 500 MRV iterations. The summary statistics consisted of positive predictive value (PPV), negative predictive value (NPV) and overall classifier accuracy. The classification was performed for comparative purposes first on original data, and afterwards on data adjusted using linear interpolation for the dosimetry applicable type of gene profile. 1,677 genes fell into the irradiation related group and 1,088 into the dosimetry applicable. The original data and adjusted data results are shown in Table 4.8.

Table 4.8: Multiple random validation metric results for analysis conducted on original expression data values values adjusted using linear interpolation of the appropriate gene profiles.

Original expression data			
	Mean [%]	Lower CI [%]	Upper CI [%]
PPV	86.71	86.13	87.29
NPV	89.32	88.76	89.89
Accuracy	87.73	87.44	88.02
Interpolation adjusted data			
PPV	93.11	92.78	93.45
NPV	94.38	94.08	94.67
Accuracy	93.56	93.39	93.72

Using a tailored approach for data classification provided a significant improvement. In the simple logistic regression model used in a multiple random validation procedure on unadjusted data, results in the context of separating control and dose-treated samples may be considered of good quality. In an alternative approach, once taking into account the gene expression profile nature of the doses applied to samples in both experiments, data in the 4 Gy timepoint were linearly interpolated to 2 Gy. However, instead of interpolating all the data non-selectively, the genes were handled according to their respective dose response profile. The adjusted data gave significantly superior results in comparison to the simple MRV scheme. The summarizing statistics excelled on adjusted data classification compared to the simple approach (positive and negative predictive value, and accuracy). This indicates that, when possible, not only increasing sample size enhances classification potential, but also using custom solutions based on knowledge of the underlying models to adjust data may be highly beneficial.

In the adjusted data, features selected for the logistic regression models in each iteration were recorded. The most frequently occurring genes were GADD45A, ZMAT3 and NAMPT. A complete list of the genes together with their occurrence frequencies is comprised in (Papiez et al., 2019).

In order to validate these findings the entire original data set was processed independently using Monte Carlo feature selection. A graphical representation of the determined interaction network was constructed and a relevant part of this network is depicted in Figure 4.15. It is clearly visible that genes involved in the highest numbers of interactions, and thus producing the largest networks, are GADD45A, ZMAT3 and CCNG1.

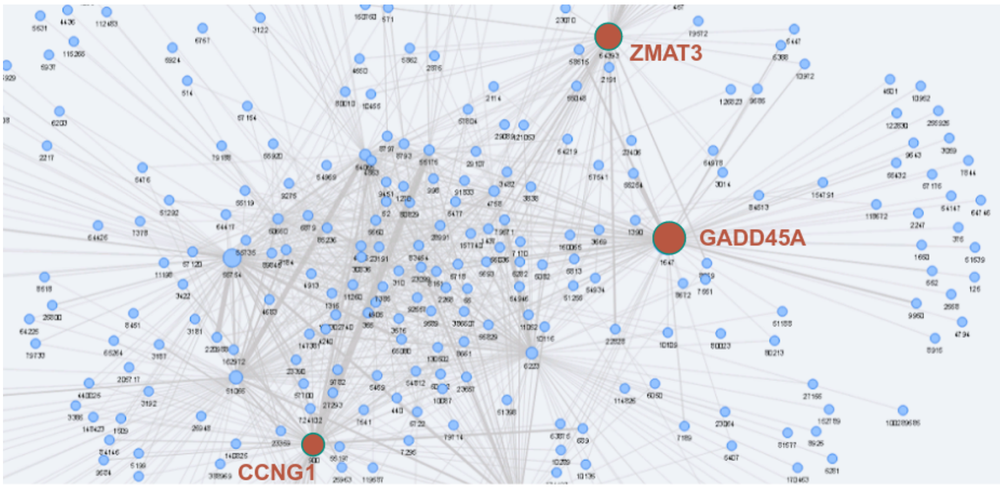


Figure 4.15: Central fragment of a gene interaction network created as an illustration of Monte Carlo feature selection results on the entire data set. The genes in bold show the highest number and largest strength of interaction with other genes.

The most often occurring gene features in the logistic regression model iterations were analyzed towards the corresponding biological function.

- GADD45A is a member of a group of genes whose transcript levels are increased following stressful growth arrest conditions and treatment with DNA-damaging agents. The DNA damage-induced transcription of this gene is mediated by both p53-dependent and -independent mechanisms. (Zhan, 2005). It has been previously proven to be a biomarker of radiation response (Kabacik et al., 2015).
- ZMAT3 mRNA and the protein are up-regulated by wildtype p53 and overexpression of this gene inhibits tumor cell growth, suggesting that this gene may have a role in the p53-dependent growth regulatory pathway (Bersani et al., 2014).
- NAMPT is thought to be involved in many important biological processes, including metabolism, stress response and aging. It has been shown to play a key

role in radiotherapy treatment(Elf et al., 2017).

Additionally, the independent feature selection method was applied and it confirmed the findings obtained by means of the MRV procedure. The MCFS method, as a rule based algorithm, focuses on genes with regard to their number and strength of interactions. The three genes, which are linked to the most interactions, also present the strongest ones (represented by width of interaction lines). These key features in case of the two merged experiments were mainly: GADD45A, ZMAT3 and CCNG1. Cyclin G1 (CCNG1) is a gene associated with G2/M phase arrest in response to DNA damage. p53 mediates its role as an inhibitor of cellular proliferation with this intermediate gene and it has previously been found to be linked with radiation response (Kabacik et al., 2015, 2011; Manning et al., 2013; Cruz-Garcia et al., 2018).

The independent identification of key features important for modeling radiation response further justifies the use of a custom data processing procedure for integrative data analysis that leads to enhanced classification. Not only have the two most significant features (GADD45A and ZMAT3) been supported with the results from a different feature selection algorithm, but also through literature research. Furthermore, this underlines the importance of investigating the less prominent genes not as single biomarkers of radiation response, but rather their impact when functioning in a network.

4.3 Multi-omics data integration

Transcriptomics and proteomics data for Mayak workers were analyzed in a combined scheme. However, this required initial customized preprocessing of the mass spectrometry data set in order to select the proteins related to dose.

4.3.1 Proteomics regression

The LC-MS/MS-based proteomics data processing identified a total of 1,281 proteins from the cardiac left ventricle samples. Dixon's outlier detection criterion was applied and no significantly outlying samples were detected in terms of protein abundance values, as the outlier distribution was uniformly spread out across the entire data set.

Strong positive correlation was determined between the factors of age and total external dose (Figure 4.16). On the contrary, no association was discovered between age or dose and body mass index (BMI). Thus, for the purpose of identifying proteins, for which abundance variation was only dose-dependent (dose is the major explanatory variable), multiple stepwise linear regression analyses were performed for each individual protein. On the other hand, proteins with only age-dependent variation were filtered, or those for which none or both of the factors (external dose and worker age) explained the existing variation. Moreover, BMI was also investigated as a factor in the regression analysis. Other clinical data, such as smoking habits and alcohol consumption could not serve as explanatory variables due to the fact that all the individuals were smokers and drinkers.

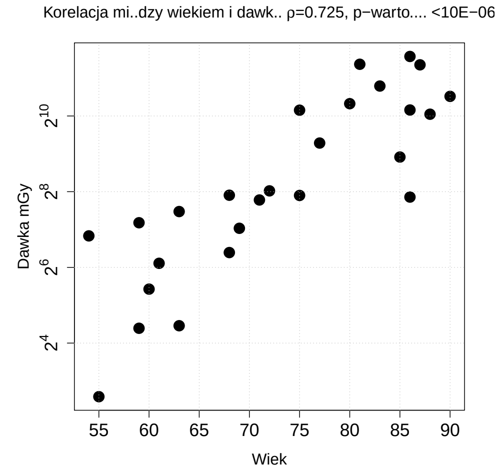


Figure 4.16: Scatter plot illustrating the data relationship between dose and age factors in the samples. Spearman's correlation coefficient with a value of 0.725 is significant (p-value $< 10e - 06$).

Altogether 582 proteins (out of 1,281) were identified as only dose-dependent (from now on the "dose-only" category), 225 as only age-dependent ("age-only"), and for 212 cases the variation was explained with a model built on the two factors. In the case of only 17 proteins, the BMI served as the dominant explanatory variable. The complete list of proteins along with the factors that constitute their respective models is available in (Papiez et al., 2018a).

Within the dose-only group of proteins, the most significant (p-value $< 10^{-7}$) were: histidine ammonia lyase (HAL), zyg-11 family member A, cell cycle regulator

(ZYG11A), RAD9-HUS1-RAD1 interacting nuclear orphan 1 (RHNO1), A-kinase anchoring protein 9 (LRG_331), moesin (MSN), acyl-CoA synthetase long chain family member 1 (ACSL1), isocitrate dehydrogenase 3 [NAD(+)] alpha (IDH3A), phosphoglycerate mutase 1 (PGAM1), chloride intracellular channel 1 (CLIC1), malic enzyme 1 (ME1), glycogen phosphorylase (PYGM), aladin WD repeat nucleoporin (AAAS), and ribosomal protein S27a (RPS27A). Multiple proteins from the above mentioned participate in processes related to energy metabolism.

The impact of the age factor was mostly weaker. Only 3 proteins, radixin (RDX), coatomer protein complex subunit beta 2 (COPB2), and protein arginine methyltransferase 5 (PRMT5) fell below significance at the level of p-value $< 10^{-5}$ in age-only models.

The dose-only and age-only dependent proteins were investigated for overrepresentation using the Gene Ontology terms repository and Kyoto Encyclopedia of Genes and Genomes pathway analyses. The most significant pathways according to GO terms were "oxidation-reduction process" and "respiratory electron transport chain" for dose-only and age-only deregulated proteins, respectively. The most significant KEGG pathways were "metabolic pathways" and "oxidative phosphorylation" for dose-only and age-only deregulated proteins, respectively. The heart-relevant enriched age-only, dose-only, and dose-age dependent KEGG pathways are presented in Table 4.9. The complete lists of age- and dose-dependent overrepresented GO and KEGG terms are available in (Papiez et al., 2018a).

Despite the strong correlation, which appeared between dose and age factors in the proteomics data, a significant majority of the differentiating proteins fell into the dose-only dependent category. The comparative analysis of the pathways activated by age-only and dose-only proteins presented general heart pathologies such as Hypertrophic cardiomyopathy and Dilated cardiomyopathy or linked to processes such as energy metabolism (Metabolic pathways, Propanoate metabolism, Oxidative phosphorylation). However, when considering dose-only related pathways the results uphold previously recorded metabolic networks (Azimzadeh et al., 2017), namely PPAR signaling, Glycolysis, Fatty acid metabolism and TCA cycle.

Table 4.9: KEGG pathways enriched by proteins that were identified as dose-only dependent and/or age-only dependent in the backward stepwise regression model selection procedure. Pathways common for dose-only and age-only dependent proteins are indicated in the bottom left.

Age-dependent	**Dose-dependent**	
Fatty acid elongation	PI3K-Akt signaling pathway	Ribosome
Tryptophan metabolism	Pathogenic Escherichia coli infection	Carbon metabolism
	Protein processing in endoplasmic reticulum	Glyoxylate and dicarboxylate metabolism
	Biosynthesis of amino acids	Arrhythmogenic right ventricular cardiomyopathy
	Proteasome	Pyruvate metabolism
Dose-age-dependent	Tight junction	Butanoate metabolism
Metabolic pathways	Glycolysis/Gluconeogenesis	Adrenergic signaling in cardiomyocytes
Cardiac muscle contraction	Peroxisome	AMPK signaling pathway
Propanoate metabolism	Leukocyte transendothelial migration	Vasopressin-regulated water reabsorption
Valine, leucine and isoleucine degradation	Fatty acid metabolism	Beta-Alanine metabolism
Hypertrophic cardiomyopathy	ECM-receptor interaction	Antigen processing and presentation
Dilated cardiomyopathy	PPAR signaling pathway	Phagosome
Oxidative phosphorylation	Fatty acid degradation	TCA cycle
	Porphyrin and chlorophyll metabolism	2-Oxocarboxylic acid metabolism
	Focal adhesion	

$\cup-$shape and $\cap-$shape protein filtration analysis

Despite of the regression analysis revealing proteins whose expression profile changed with dose, it does not take into consideration cases when changes show no linear relation of the dose. This situation may be designated as a $\cup-$shape or a $\cap-$shape. The former exists when the protein levels are high in the control group, then decline with dose, and then increase again along with the doses. The latter occurs where in the control group the protein level is low, increases with dose, and at the highest doses decreases. An example is presented in Figure 4.17.

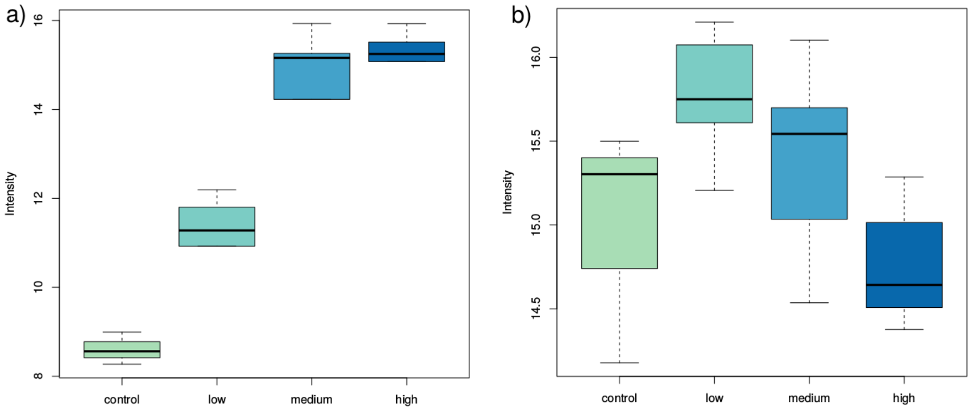


Figure 4.17: Example boxplots for two proteins identified in the regression analysis as dose-only dependent. The left plot represents a situation where the protein level gradually grows, whereas the right the protein level forms a ∩−shape with an increase in the low-dose groups and a decrease in the higher doses.

In order to enable analysis of such situations and acquire a comprehensive description of the deregulation with filtered out ∪−shape or a ∩−shape proteins, tests for protein abundance differentiation were performed on the proteomics samples (Azimzadeh et al., 2017). Subgroups of low- (< 100 mGy), medium- ($100 - 500$ mGy) and high-dose (> 500 mGy) exposed samples were all analyzed with regard to the non-exposed control group. Therefore, The Kruskal-Wallis test for differentiation between the dose groups was carried out, as Lilliefors test results proved the normality assumption not fulfilled in the subgroups. It also confirmed the regression analysis results (582 dose-only dependent proteins), as 582 proteins were found to be significantly deregulated between at least one of the four dose groups (control, < 100 mGy, $100 - 500$ mGy, and > 500 mGy). Post-hoc Dunnett tests enabled the discovery of sets of deregulated proteins specific for every external dose group (Table 4.10). Even though for many proteins their expression increased with the dose, an overwhelming majority of all deregulated proteins were identified as down-regulated.

The numbers of total and intersecting differentiating proteins in the dose groups are illustrated on Venn diagrams in Figure 4.18.

When examining the group of up-regulated proteins, twelve were significantly changed when compared to unexposed samples in both high and medium external dose groups. Only one uncharacterized previously protein (C1orf112) was common for

Table 4.10: Numbers of dose-only dependent deregulated proteins in different dose groups in comparison to the non-exposed controls resulting from post-hoc Dunnett tests.

With reference to controls	$<$ 100 mGy	$100 - 500$ mGy	$>$ 500 mGy
Up-regulated proteins	1	15	12
Down-regulated proteins	2	33	307
$\cup-$**shape and** $\cap-$**shape proteins**	260		

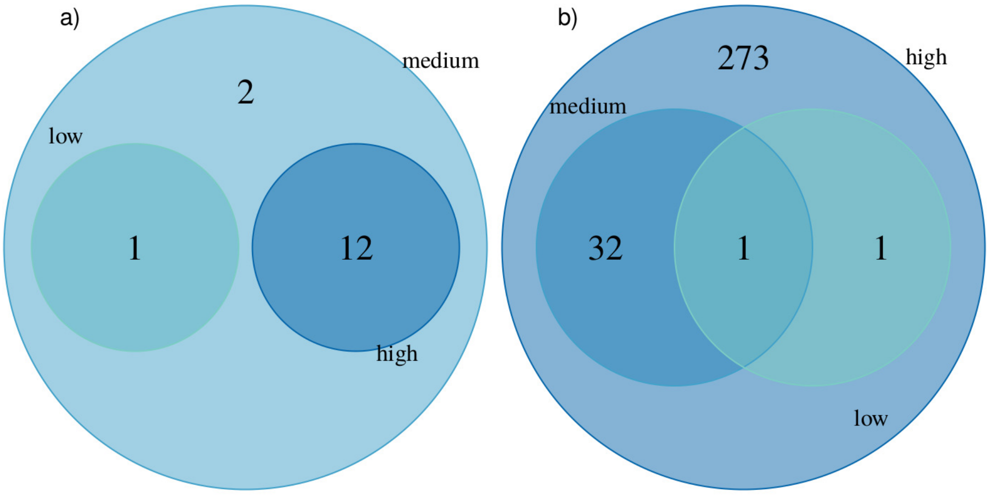


Figure 4.18: Venn diagrams presenting the numbers and overlap of significantly **a)** up-regulated and **b)** down-regulated proteins in different dose groups among dose-only dependent proteins with respect to the control according to Dunnett's test at $\alpha = 0.05$.

up-regulated proteins in low and medium external dose individuals but its abundance among samples of the high-dose group did not differ significantly from the unexposed samples. None of the proteins were significantly up-regulated in all dose groups.

In the down-regulated proteins, the cytochrome C oxidase assembly factor (COX20) protein demonstrated significant deregulation in all dose groups (low, medium and high) in relation to the control samples. This is of importance, as it may potentially represent a switch-type biomarker of radiation exposure (Figure 4.20). This protein plays a key role in the assembly of cytochrome C oxidase, an essential component of the respiratory pathway.

Moreover, the 32 common differentially regulated proteins between the medium and high doses are: ALAD, ARHGAP11A, ATP5L, BFSP1, C14orf2, CA2, CCDC141, CRAT, DLD, EIF2B5, FECH, FNDC3A, HSD17B4, ITGA6, LAP3, LGALS3BP, LRG_391, LRRC37B, MCCC1, MEMO1, MLYCD, MYOM3, NDUFB11, OLA1, OTUB1, PCBD2, RAB5A, RXRA, SLC25A3, SUCLA2, UCHL3, WIPI1). Twelve of these proteins are located in the mitochondria and/or have metabolic functions. Most of the down-regulated proteins (273) appear only in the high dose group, which suggests that their expression is not linearly dose-dependent, but rather by a high dose threshold.

4.3.2 Analysis of differentially regulated transcripts

In the course of RNA-seq analysis 25,221 transcripts were identified for the 4-sample data set comprising two control and two high-dose samples. The DEseq adaptive threshold method for filtering low count data led to eliminating transcripts with 4 or less counts mapped from onward analysis. The negative binomial test used to determine differentially expressed transcripts with Benjamini-Hochberg multiple testing correction provided comparable numbers of 979 significantly up-regulated transcripts and 895 significantly down-regulated transcripts.

4.3.3 Proteomics and transcriptomics integration

For the purpose of combining the proteomics data with the RNA-seq data, only the high-dose samples group could be taken into further investigation to the integration procedure. Firstly, after applying hierarchical clustering in the transcriptomics data, the high-dose samples could be clearly separated from the controls. Similarly, the dose-only proteins successfully separated high-dose and control samples (Figure 4.19). By contrast, the supervised clustering analysis with age-related protein features did not produce a heatmap where the individuals would be grouped by age in a consistent manner but still rather by the respective doses. The full lists of differentially expressed genes and deregulated proteins along with the corresponding p-values are compiled in (Papiez et al., 2018a).

When considering only the common differentiating protein and transcript pairs at the level of 5% in both data sets, (this approach is henceforth defined as the restrictive), only 2 protein-transcript pairs (ANK3, P4HTM) overlapped as statistically significantly up-regulated and 30 as down-regulated (ACADM, ANXA1, ANXA5, CALM2, CAP1, CD93, DCN, DLD, DPT, DSTN, EIF4A2, ERAP1, GLRX, GRPEL1, HNRNPK, HSPA8, ITGA6, LAP3, LGALS1, LUM, NIPSNAP3A, NIPSNAP3B, PDIA3, RAB5A, RBBP7, RPS4X, RPS6, SDPR, SUCLG1, UBE2N).

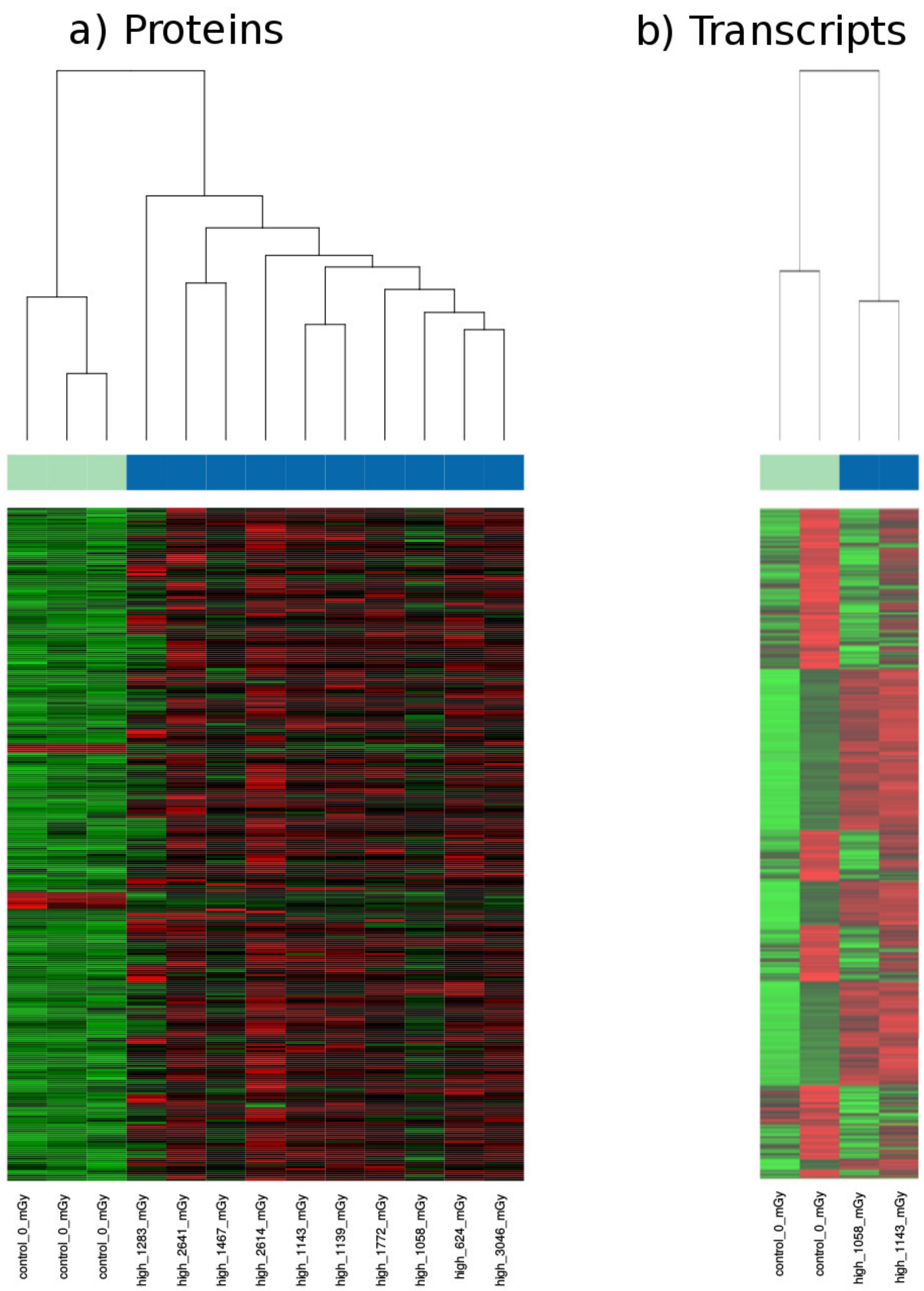


Figure 4.19: Supervised heat map showing the separation of high-dose samples from controls based on **a)** 319 dose-dependent significantly deregulated proteins; and **b)** 1,874 significantly deregulated transcripts. The numbers provided next to the class label show total external dose of the individual. The color bars indicate sample groups: cyan - controls, blue - high-dose samples.

Among the down-regulated gene/protein pairs, several were members of the following molecular function GO terms: RNA binding, Oxidoreductase activity, and Poly(A) RNA binding. It was a condition for the transcript/protein pairs to be coherent in terms of the direction of the deregulation in order to include them into the analysis, as for instance shown in Figure 4.20.

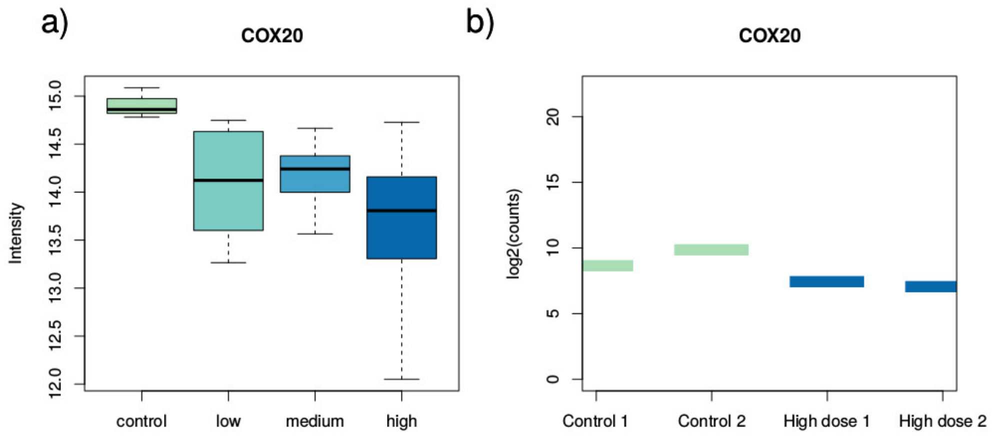


Figure 4.20: The down-regulated COX20 shown as an exemplary coherent deregulated transcript-protein pair. The boxplots in **a)** show the statistical summary of protein expression values in each dose group, whereas the bars in plot **b)** present the gene expression for each available individual transcript sample on the logarithmic scale. The plot illustrates the downward trend in the direction of protein expression, and though there is a large range of values in the high-dose group distribution, a clearly significant difference is be observed compared to the control group. Likewise, in the RNA-seq data downregulation is noticeable in this case.

As previously discussed, when applying the restrictive approach only shared significantly deregulated transcripts and proteins from the two data sets are taken into account. Nevertheless, the binary decision of identifying a transcript/protein as deregulated solely considering whether it falls below a fixed significance threshold is often the cause of excluding valid results and applying an adaptive approach would be in most cases more favorable.

Thus, the integrative approach was implemented to limit the chance of discarding important information, by considering the actual strength of differentiation expressed through p-values. Therefore, Fisher's statistical integration was employed on the complete sets of p-values from the negative binomial test for transcripts and Dunnett's test for high-dose proteins. With a combined p-value threshold of 0.05, additional 363 transcript-protein pairs were identified as significantly deregulated in the integrative approach (from a total of 395: 32 in the restrictive approach and additional 363 in the integrative). After Benjamini-Hochberg multiple testing correction, 69 transcripts prevailed as significant for the p-value integration. The significant transcript-protein pairs along with the corresponding p-values are listed in (Papiez et al., 2018a).

The deregulated features identified using the restrictive approach in proteomics

(319 dose-only dependent proteins deregulated between control and high dose) and transcriptomics (1,874 differentiating transcripts) were examined for enrichment and common GO terms. Furthermore, relevant KEGG signaling pathways were investigated. analogous overrepresentation analysis was also carried out in the case of transcripts validated by proteins in the integrative approach (69 coherent transcript-protein pairs). The overrepresented KEGG pathways with respect to the applied approach are presented in Table 4.11. Only one KEGG pathway, Propanoate metabolism, was commonly enriched in the two approaches. A detailed list of overrepresented pathways with the corresponding proteins is available in (Papiez et al., 2018a).

Table 4.11: KEGG signaling pathways overrepresented by gene-protein pairs found to be significantly deregulated in high-dose samples in comparison to controls. The pathways in the left column were obtained from the intersection of enriched pathways from significant genes and proteins in the two data sets. The pathways in the right column were enriched by gene-protein features significant by the combined Fisher's p-value method.

Restrictive approach	**Integrative approach**	
Proteasome	Glycolysis / Gluconeogenesis	Beta-Alanine metabolism
Ribosome	Oxidative phosphorylation	Metabolic pathways
Proteoglycans in cancer	Citrate cycle (TCA cycle)	Tryptophan metabolism
Pathogenic Escherichia coli infection	Bacterial invasion of epithelial cells	Arginine and proline metabolism
Propanoate metabolism		Lysine degradation
	Phagosome	PPAR signaling pathway
	Vasopressin-regulated water reabsorption	Proximal tubule bicarbonate reclamation
	Ascorbate and aldarate metabolism	Terpenoid backbone biosynthesis
	Valine, leucine and isoleucine degradation	Glyoxylate and dicarboxylate metabolism
	Histidine metabolism	Fatty acid degradation
	Pyruvate metabolism	Carbon metabolism

In total overrepresented terms constituted 241 GO Biological Process ontologies discovered with the restrictive approach and 54 identified in the integrative approach. 24 of the enriched terms were common between the two methods.The full list of overrepresented ontologies is available in (Papiez et al., 2018a).

In conclusion, the integrative approach was determined as superior over the restrictive comparison in validating the proteomics data results using the transcriptomics data analysis. The significance of Fisher's combined p-value gene-protein pairs provided links to overrepresented KEGG pathways, which were chiefly radiation-linked processes, e.g. PPAR signaling, TCA cycle and Glycolysis/Gluconeogenesis. By contrast the common of KEGG terms overrepresented in the separately analyzed proteomics and transcriptomics data sets were few and not specific, including notions such as Proteasome, or Ribosome. These do represent two main cellular machineries highly dependent on energy supply for cellular functions and include proteins important in oxidoreductase activity (Proteasome) and RNA binding proteins (Ribosome), yet the more distinguishing processes would not be discovered without the use of the integrative data analysis workflow.

4.4 Inter-tissue data integration

The total number of proteins identified in the tissues was 831 in the fibroblast cells, 791 in coronary artery, 332 in mammary cells, and 1097 in leukocytes. The overlap between the proteins is presented in Figure 4.21.

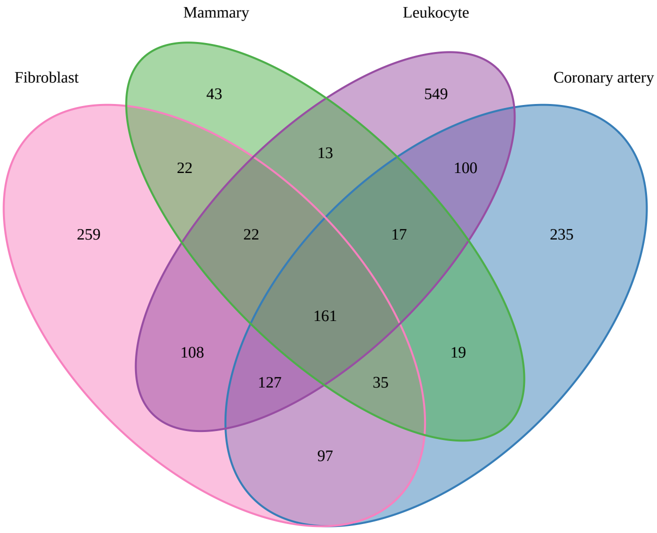


Figure 4.21: Numbers of proteins common between the four cell systems.

The samples were analyzed in terms of similarity among tissues using the 161 proteins identified across all of the cell systems. Hierarchical clustering was carried out based on the 161 proteins and the resulting heatmap in Figure 4.22 clearly illustrates that tissue is the dominant factor differentiating samples.

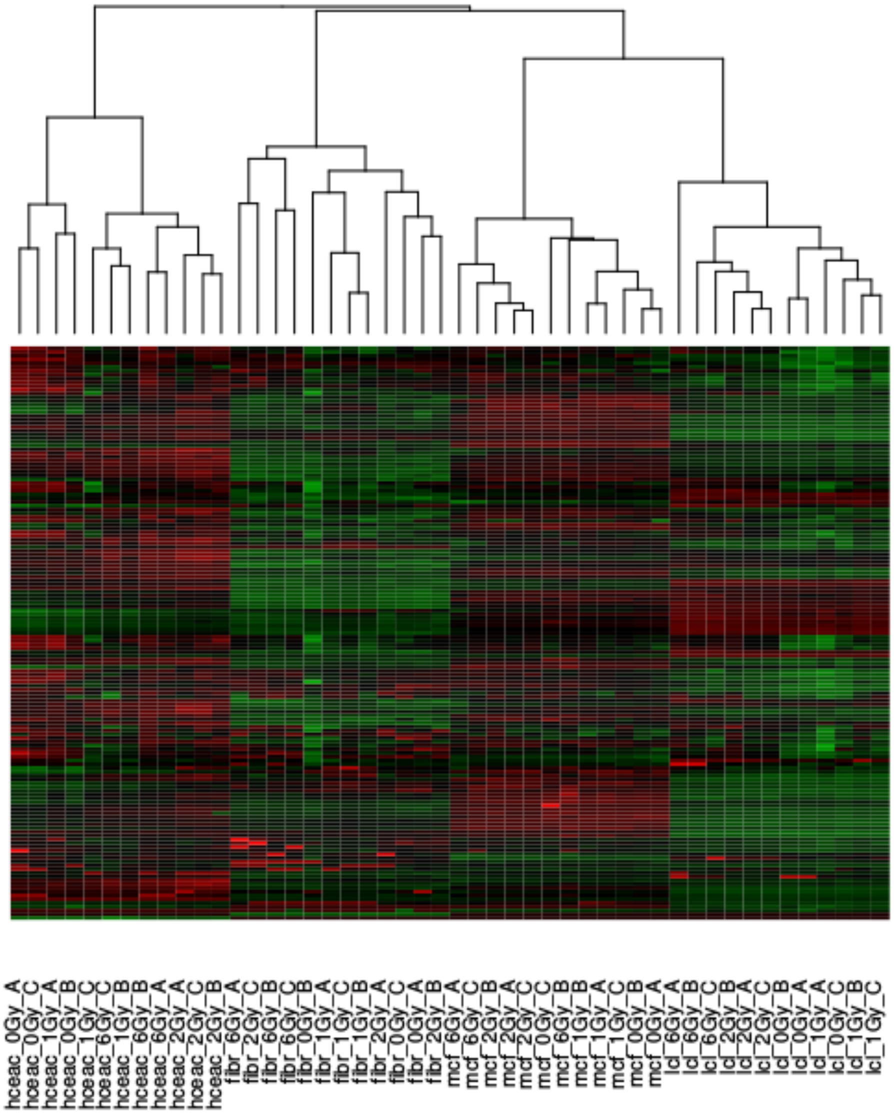
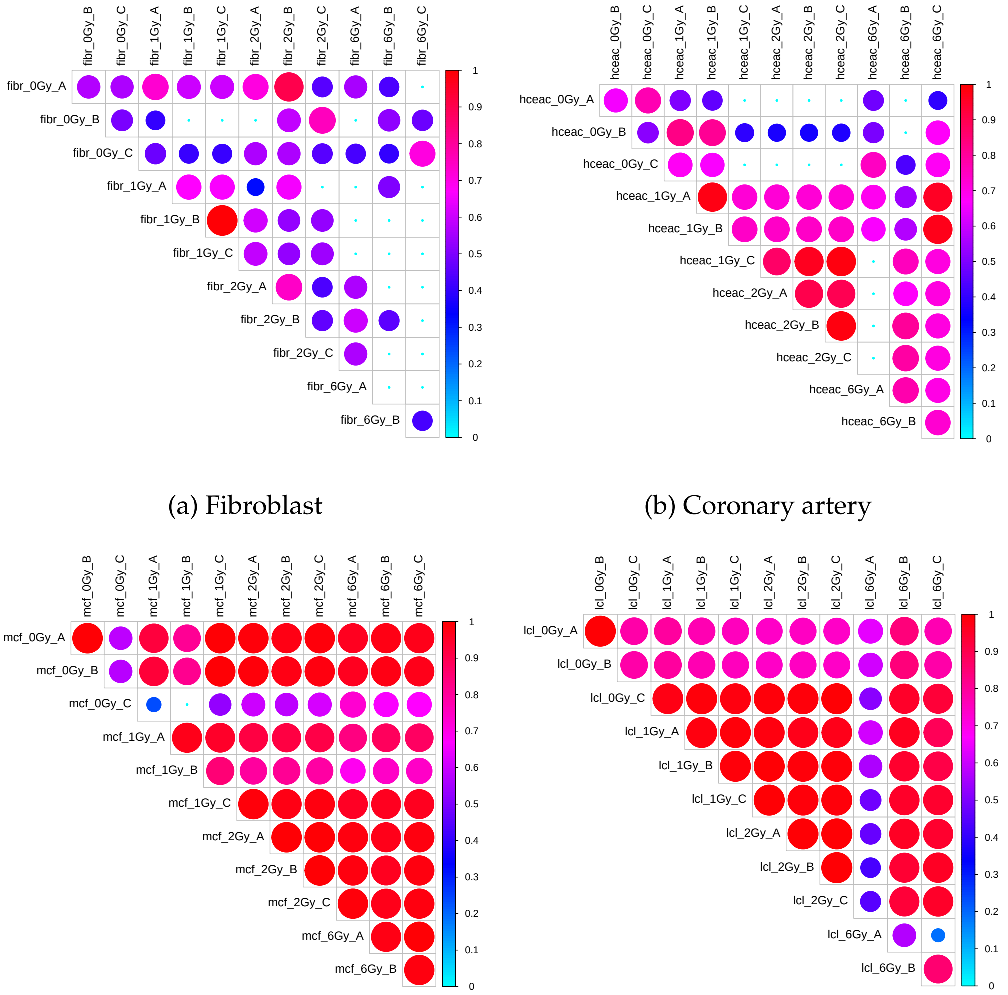


Figure 4.22: Heatmap illustrating clustering between exosome samples. Sample names are built in the pattern: tissue_dose_replicate. The dominant factor differentiating the samples is unequivocally the cell system.

Following the clustering analysis, similarity metrics were calculated within the tissue groups between individual samples (Figure 4.23). The results indicate that there are two outliers in the Fibroblast tissue that show relatively high similarity to mammary samples (Fibr_0Gy_C and Fibr_6Gy_C). Furthermore, within the mammary and leukocyte tissue groups samples show high similarity regardless of the dose group. In coronary artery samples the similarity score is relatively high within the dose groups.

The lowest similarity may be observed among some of the fibroblast samples, especially within the 6 Gy dose group.



(a) Fibroblast

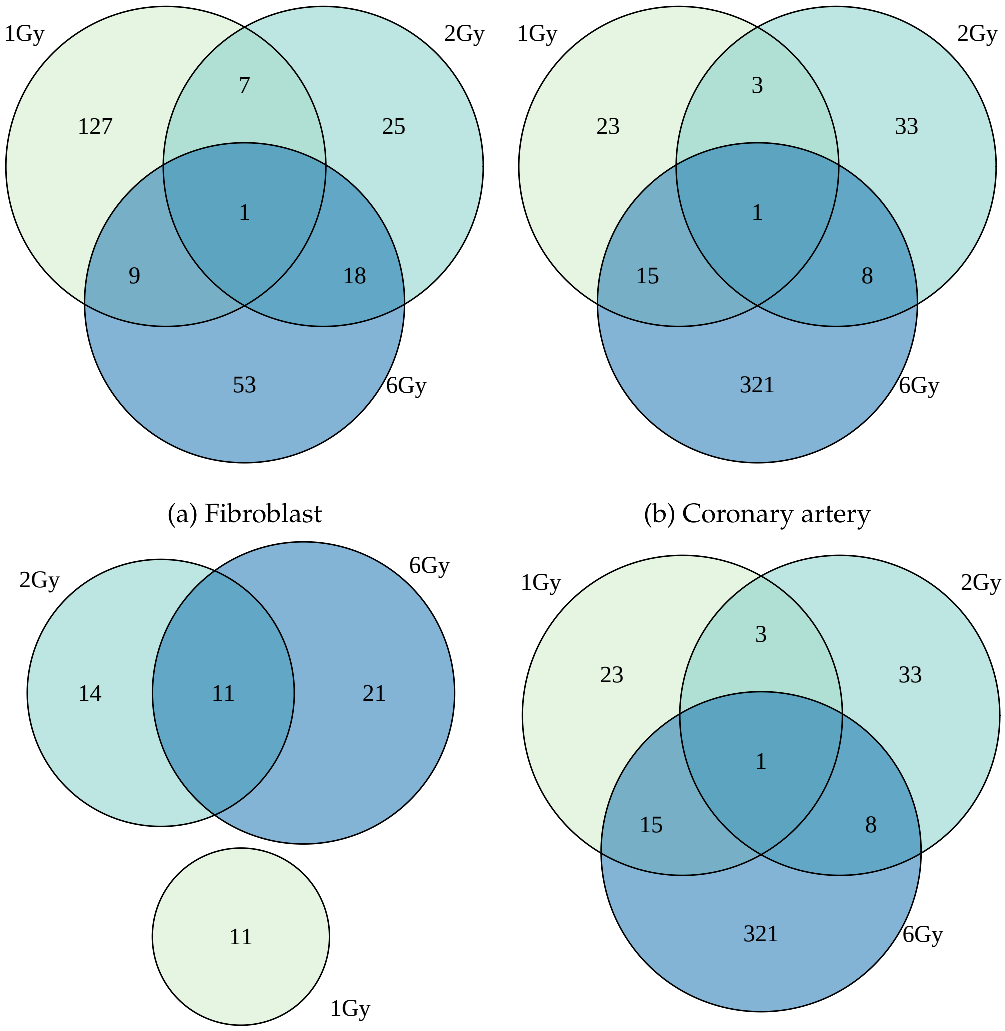
(b) Coronary artery

(c) Mammary

(d) Leukocyte

Figure 4.23: Similarity graphs for individual tissue groups. The size of the circle reflects values of the similarity metric, as well as the color bar.

Dunnett's test was performed in order to find proteins differentiating between consecutive doses and controls in entire separate data sets. Within nearly all of the tissue groups the largest numbers of overlapping significantly deregulated proteins are present between the 2 Gy and 6 Gy dose groups. Only in the case of leukocyte samples there are more common proteins between the 1 Gy and 6 Gy dose (Figure 4.24). Hierarchical clustering based on deregulated proteins shows clear separation of the dose groups in the fibroblast, coronary artery and leukocyte tissues. In case of the mammary tissue, the samples are mixed among the dose groups. In fibroblasts, the controls are the most isolated group of samples. In coronary artery endothelial cells, the 6 Gy dose samples are grouped closer together with the 1 Gy dose samples than 2 Gy dose (Figure 4.25). This combined analysis may serve as a basis for the design of experiments that will provide a deeper insight into the different cell types common patterns of radiation response. Moreover, the question of a common cargo of exosomes from different cell types is a point of interest for further investigation.



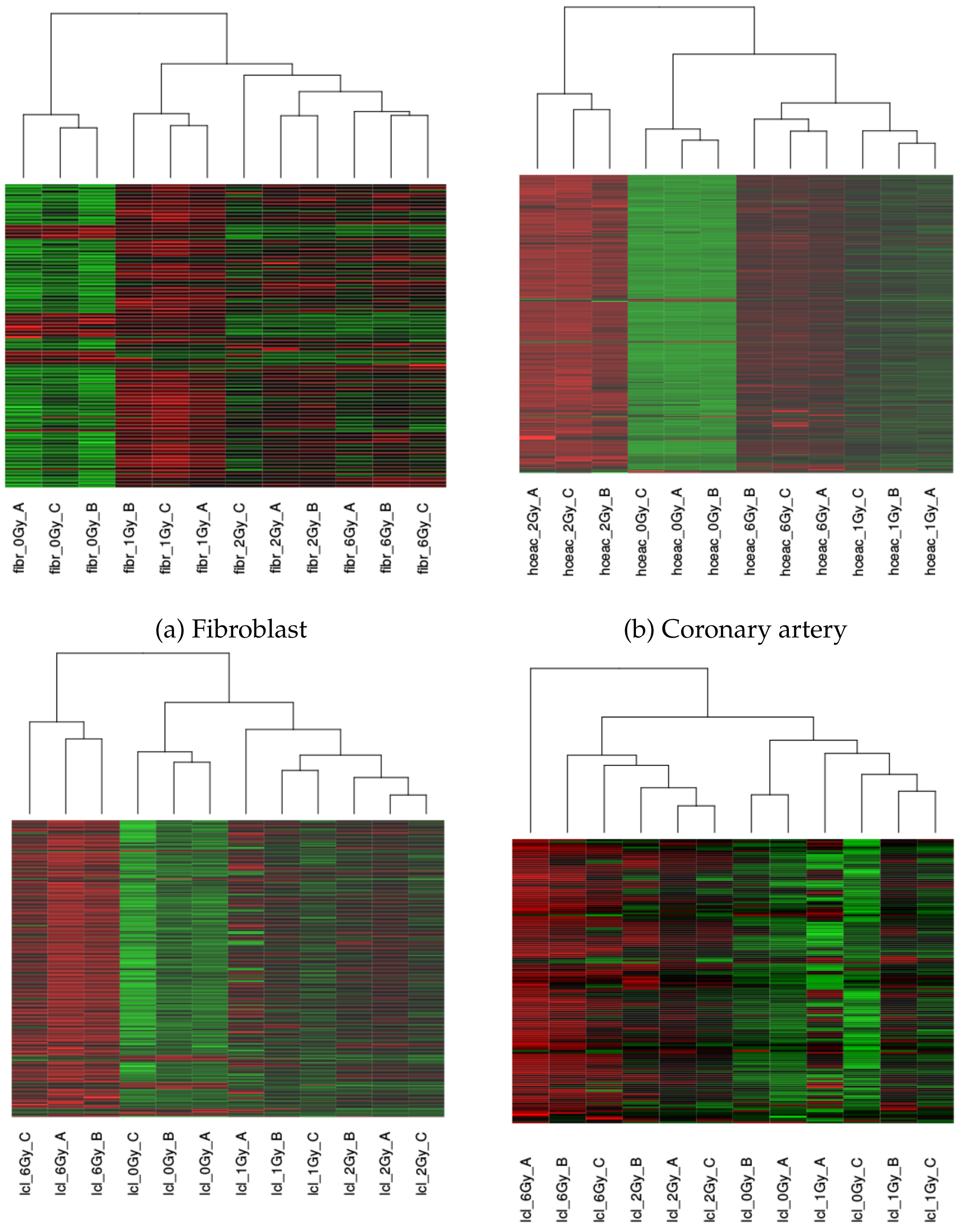
(a) Fibroblast

(b) Coronary artery

(c) Mammary

(d) Leukocyte

Figure 4.24: Deregulated protein Venn diagrams for individual tissue groups based on Dunnett's of dose group against controls.



(a) Fibroblast

(b) Coronary artery

(c) Mammary

(d) Leukocyte

Figure 4.25: Heatmaps of samples in tissue groups based on deregulated proteins.

Chapter 5

Conclusions

The objective of this thesis in unraveling the potential of statistical and data mining integration techniques for the purpose of biomarker disease research has been achieved in a number of various aspects.

5.1 Integration within an experiment

Firstly, it has been demonstrated that the identification and correction of batch effects is an indispensable stage in high-throughput molecular biology data preprocessing. In this work an efficient and unique method of batch effect identification has been proposed. It enables the partitioning of data into corresponding batches before processing with the use of correction tools, which require prior knowledge of batch structure. The identification is carried out by means of a dynamic programming approach and batch number selection is conducted with the δ gPCA statistic.

The algorithm's performance on recovering previously known batch divisions was proved to be highly efficient when considering the four assessed experiments (microarray, RNA-seq, mass spectrometry) with the use of average Dice Index as a similarity measure. Furthermore, when analyzing sets where batch structure was not given *a priori*, intra-group correlation showed that in a majority of cases data integrity increases within groups formed by the studied biological processes (case/control) after correction of the identified batch effects. This was further highlighted by provoking a significant change in the proportion of total variance present in the data explained by batch effects.

Additionally, literature and Gene Ontology term referencing implies that adequate and tailored batch effect preprocessing entails potential new discoveries of knowledge relevant to the studied disease or condition. This was further emphasized with an observed increase in functional Information Content. By contrast, the failure to take into consideration batch effects, when their share in the total variation is large, may lead to insignificant conclusions and impede the development of a studied disease, by omission of important conclusions, which may be drawn from the performed experiments.

5.2 Integration within an omics

Combining research of high-throughput transcriptomic data constitutes a relevant solution to the problem of dimensionality reduction, yet they impose a challenge in terms of transformation of the measurements to achieve computational and biological consistency. This issue becomes more complex when the compared data set platform design varies. Therefore, a procedure for integrated study of data from oligonucleotide and cDNA microarrays was established, to facilitate the merging of expression sets and making them comparable both in the numerical form and within the analyzed biological condition. The limitations of this method include loss of information about features unique for either of the platforms. However, due to the raise of statistical power after merging of the two data sets, enhanced results were obtained in the form of an increase of information about differentially expressed genes and the additional features have been shown to be annotated to processes related to the studied question of radiosensitivity.

The proposed strategy for gene signature selection involving statistical integration of p-values provided supreme results when considering radioresistant and radiosensitive patient separability. This was demonstrated using two classifier models, namely logistic regression and support vector machine. The integrative approach provided both a decline in error rates and upturns in positive and negative predictive values. Additionally, the obtained optimal gene signature presents links to radiosensitivity and cancer-related processes occurring in relevant signaling pathways. These findings imply the use of gene signatures obtained through integrative methods in more complex classification problems such as multiple random validation.

Therefore, a customized approach for high-throughput transcriptomic data analysis was tested, based on statistical integration tools and current knowledge of biological mechanisms. Gene profiles were applied as a filtering factor to adjust data using linear interpolation to allow for efficient classification in a multiple random validation setting. The implementation of integrative techniques combined with custom data interpolation between doses led to successful determination of potential biomarkers of radiation response, which have been confirmed with an independent computational approach (MCFS) and literature study.

To summarize, *in silico* machine learning analysis combined with integrative statistical techniques with functional validation and profile modeling formed a comprehensive solution for the discovery of dose response mechanisms and revealing features, which are the most applicable to form a signature. The idea of using tailored procedures involving data integration narrows down the search area for experts, potentially saving time and effort and allowing for improvement in planing the design of future biological experiments held with the purpose of studying diseases treatment mechanisms.

5.3 Inter-omics and inter-tissue integration

In the transcriptomics and proteomics data analysis for elucidating mechanisms of radiation-induced ischemic heart disease a substantial contribution was achieved with the use of custom statistical methods to distinguish dose-only dependent protein expression changes from the age-only dependent changes. On top of that, the use of an integrative statistical analysis approach, adapted to the nature of the studied data served as an alternative validation procedure for the discovered proteomic processes by the gene expression study. Pathways such as glycolysis, oxidative phosphorylation, citric acid cycle and, importantly, PPAR signaling were confirmed using the p-value integration technique. Verification of the gained knowledge was not possible when applying a conventional restrictive result comparison procedure. This outcome emphasizes the importance of careful planning and the benefits of non-standard data set merging pipelines for maximizing the chance of obtaining valid conclusions.

The example of multi-tissue analysis of exosome proteomics data exposes the importance of undertaking complex tasks for the purpose of explaining biological processes in a comprehensive manner. The similarity study between four different cell systems demonstrated factors that play a key role in the differentiation of irradiated samples. Moreover, the qualitative differences among tissues form a starting point for designing additional experiments for exosome proteomic pattern investigation.

The algorithms and methods proposed in this work constitute an entirely original contribution in the field of statistical and data mining analyses of high-throughput molecular biology data sets. The dynamic-programming batch identification method is a novel tool, available to the scientific community through the BatchI R package implementation. Likewise, the integration procedures at the level of experimental platforms, omics and tissues have not been compiled and applied in such applications previously, and moreover with the auspicious results provided, they may serve as a base for improvements in the continuous efforts towards the development of biomedical data analysis techniques of the future.

Bibliography

Abbott, A. et al. (2015). Researchers pin down risks of low-dose radiation. *Nature*, 523(7558):17–8.

Aebersold, R. and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, 422(6928):198.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Alexa, A. and Rahnenfuhrer, J. (2010). topGO: enrichment analysis for gene ontology. *R package version*, 2(30).

Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106.

Antoniadis, A., Lambert-Lacroix, S., and Leblanc, F. (2003). Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics*, 19(5):563–570.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25.

Auer, P. L. and Doerge, R. (2010). Statistical design and analysis of RNA sequencing data. *Genetics*, 185(2):405–416.

Azimzadeh, O., Azizova, T., Merl-Pham, J., Subramanian, V., Bakshi, M. V., Moseeva, M., Zubkova, O., Hauck, S. M., Anastasov, N., Atkinson, M. J., et al. (2017). A dose-dependent perturbation in cardiac energy metabolism is linked to radiation-induced ischemic heart disease in Mayak nuclear workers. *Oncotarget*, 8(6):9067.

Baggerly, K. A., Coombes, K. R., and Morris, J. S. (2006). An introduction to high-throughput bioinformatics data. *Bayesian Inference for Gene Expression and Proteomics*, 1:1–39.

Bakker, O. B., Aguirre-Gamboa, R., Sanna, S., Oosting, M., Smeekens, S. P., Jaeger, M., Zorro, M., Võsa, U., Withoff, S., Netea-Maier, R. T., et al. (2018). Integration of multi-omics data and deep phenotyping enables prediction of cytokine responses. *Nature immunology*, page 1.

Bellman, R. (1961). On the approximation of curves by line segments using dynamic programming. *Communications of the ACM*, 4(6):284.

Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M., and Marron, J. S. (2004). Adjustment of systematic microarray data biases. *Bioinformatics*, 20(1):105–114.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122.

Bersani, C., Xu, L., Vilborg, A., Lui, W., and Wiman, K. (2014). Wig-1 regulates cell cycle arrest and cell death through the p53 targets FAS and 14-3-3σ. *Oncogene*, 33(35):4407.

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

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breitling, R., Armengaud, P., Amtmann, A., and Herzyk, P. (2004). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS letters*, 573(1):83–92.

Brodsky, R. A., Vala, M. S., Barber, J. P., Medof, M. E., and Jones, R. J. (1997). Resistance to apoptosis caused by PIG-A gene mutations in paroxysmal nocturnal hemoglobinuria. *Proceedings of the National Academy of Sciences*, 94(16):8756–8760.

Bylesjö, M., Eriksson, D., Sjödin, A., Jansson, S., Moritz, T., and Trygg, J. (2007). Orthogonal projections to latent structures as a strategy for microarray data normalization. *BMC bioinformatics*, 8(1):207.

Chaudhary, K., Poirion, O. B., Lu, L., and Garmire, L. X. (2018). Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clinical Cancer Research*, 24(6):1248–1259.

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., and Liu, C. (2011). Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PloS one*, 6(2):e17238.

Chen, X., Teichmann, S. A., and Meyer, K. B. (2018). From tissues to cell types and back: Single-cell gene expression analysis of tissue architecture. *Annual Review of Biomedical Data Science*, 1:29–51.

Chung, E. J., Brown, A. P., Asano, H., Mandler, M., Burgan, W. E., Carter, D., Camphausen, K., and Citrin, D. (2009). In vitro and in vivo radiosensitization with AZD6244 (ARRY-142886), an inhibitor of mitogen-activated protein kinase/extracellular signal-regulated kinase 1/2 kinase. *Clin. Cancer Res.*, 15(9):3050–3057.

Cruz-Garcia, L., O'Brien, G., Donovan, E., Gothard, L., Boyle, S., Laval, A., Testard, I., Ponge, L., Woźniak, G., et al. (2018). Influence of confounding factors on radiation dose estimation in in vivo validated transcriptional biomarkers. *Health Physics*.

Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., Bunney, W. E., Myers, R. M., Speed, T. P., Akil, H., et al. (2005). Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic acids research*, 33(20):e175–e175.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Dimitrakopoulos, C., Hindupur, S. K., Häfliger, L., Behr, J., Montazeri, H., Hall, M. N., and Beerenwinkel, N. (2018). Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics*, 34(14):2441–2448.

Ding, M., Zhang, E., He, R., and Wang, X. (2013). Newly developed strategies for improving sensitivity to radiation by targeting signal pathways in cancer therapy. *Cancer science*, 104(11):1401–1410.

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

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121.

Edgar, J. R. (2016). Q&a: What are exosomes, exactly? *BMC biology*, 14(1):46.

Elf, A.-K., Bernhardt, P., Hofving, T., Arvidsson, Y., Forssell-Aronsson, E., Wängberg, B., Nilsson, O., and Johanson, V. (2017). NAMPT inhibitor GMX1778 enhances the efficacy of 177Lu-DOTATATE treatment of neuroendocrine tumors. *Journal of Nuclear Medicine*, 58(2):288–292.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D., and Bray, F. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer*, 136(5):E359–E386.

Finnon, P., Kabacik, S., MacKay, A., Raffy, C., A'Hern, R., Owen, R., Badie, C., Yarnold, J., and Bouffler, S. (2012). Correlation of in vitro lymphocyte radiosensitivity and gene expression with late normal tissue reactions following curative radiotherapy for breast cancer. *Radiotherapy and oncology*, 105(3):329–336.

Fisher, R. A. (1992). Statistical methods for research workers. In *Breakthroughs in Statistics*, pages 66–70. Springer.

Fisher, R. A. (2006). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

Fox, D. A., Chiorazzi, N., and Katz, D. H. (1976). Hapten specific ige antibody responses in mice v. differential resistance of ige and igg b lymphocytes to x-irradiation. *The Journal of Immunology*, 117(5 Part 1):1622–1628.

Frank, A. M., Bandeira, N., Shen, Z., Tanner, S., Briggs, S. P., Smith, R. D., and Pevzner, P. A. (2007). Clustering millions of tandem mass spectra. *Journal of proteome research*, 7(01):113–122.

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

Gajula, M. P. (2016). Multi-omics data integration: A modular approach. *J Mol Genet Med*, 10(4):232.

Galamb, O., Györffy, B., Sipos, F., Spisák, S., Németh, A. M., Miheller, P., Tulassay, Z., Dinya, E., and Molnár, B. (2008). Inflammation, adenoma and cancer: objective classification of colon biopsy specimens with gene expression signature. *Disease Markers*, 25(1):1–16.

Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. (2012). Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Molecular & Cellular Proteomics*, pages mcp–M111.

Giordano, T. J., Kuick, R., Else, T., Gauger, P. G., Vinco, M., Bauersfeld, J., Sanders, D., Thomas, D. G., Doherty, G., and Hammer, G. (2009). Molecular classification and prognostication of adrenocortical tumors by transcriptome profiling. *Clinical Cancer Research*, 15(2):668–676.

Glaab, E., Garibaldi, J. M., and Krasnogor, N. (2009). Arraymining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization. *BMC bioinformatics*, 10(1):358.

Goossens, N., Nakagawa, S., Sun, X., and Hoshida, Y. (2015). Cancer biomarker discovery and validation. *Translational cancer research*, 4(3):256.

Govindarajan, R., Duraiyan, J., Kaliyappan, K., and Palanisamy, M. (2012). Microarray and its applications. *Journal of pharmacy & bioallied sciences*, 4(Suppl 2):S310.

Haggar, F. A., Boushey, R. P., et al. (2009). Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clinics in colon and rectal surgery*, 22(4):191.

Hall, M. A. (1999). Feature selection for discrete and numeric class machine learning. *Working Paper*.

Han, X., Aslanian, A., and Yates III, J. R. (2008). Mass spectrometry for proteomics. *Current opinion in chemical biology*, 12(5):483–490.

Heller, M. J. (2002). Dna microarray technology: devices, systems, and applications. *Annual review of biomedical engineering*, 4(1):129–153.

Heller, R. A., Schena, M., Chai, A., Shalon, D., Bedilion, T., Gilmore, J., Woolley, D. E., and Davis, R. W. (1997). Discovery and analysis of inflammatory disease-related genes using cdna microarrays. *Proceedings of the National Academy of Sciences*, 94(6):2150–2155.

Herok, R., Konopacka, M., Polanska, J., Swierniak, A., Rogolinski, J., Jaksik, R., Hancock, R., and Rzeszowska-Wolny, J. (2010). Bystander effects induced by medium from irradiated cells: similar transcriptome responses in irradiated and bystander k562 cells. *International Journal of Radiation Oncology* Biology* Physics*, 77(1):244–252.

Howe, G. M. and Loraine, J. A. (2013). *Environmental medicine*, chapter Diseases of Affluence. Elsevier.

Huang, S., Chaudhary, K., and Garmire, L. X. (2017). More is better: recent progress in multi-omics data integration methods. *Frontiers in genetics*, 8:84.

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

Jackson, B., Scargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., and Tsai, T. T. (2005). An algorithm for optimal partitioning of data on an interval. *Signal Processing Letters, IEEE*, 12(2):105–108.

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

Jonckheere, A. R. (1954). A distribution-free k-sample test against ordered alternatives. *Biometrika*, 41(1/2):133–145.

Kabacik, S., Mackay, A., Tamber, N., Manning, G., Finnon, P., Paillier, F., Ashworth, A., Bouffler, S., and Badie, C. (2011). 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.

Kabacik, S., Manning, G., Raffy, C., Bouffler, S., and Badie, C. (2015). Time, dose and ataxia telangiectasia mutated (ATM) status dependency of coding and noncoding RNA expression after ionizing radiation exposure. *Radiation research*, 183(3):325–337.

Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.

Kerkhofs, T. M., Verhoeven, R. H., Van der Zwan, J. M., Dieleman, J., Kerstens, M. N., Links, T. P., Van de Poll-Franse, L. V., and Haak, H. R. (2013). Adrenocortical carcinoma: a population-based study on incidence and survival in the Netherlands since 1993. *European Journal of Cancer*, 49(11):2579–2586.

Kohl, M., Megger, D. A., Trippler, M., Meckel, H., Ahrens, M., Bracht, T., Weber, F., Hoffmann, A.-C., Baba, H. A., Sitek, B., et al. (2014). A practical data processing workflow for multi-omics projects. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1844(1):52–62.

Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y. A., Williams, E., Dylag, M., Kurbatova, N., Brandizi, M., Burdett, T., et al. (2014). ArrayExpress update - simplifying data submissions. *Nucleic Acids Research*, page gku1057.

Kong, X., Liu, N., and Xu, X. (2014). Bioinformatics analysis of biomarkers and transcriptional factor motifs in down syndrome. *Brazilian Journal of Medical and Biological Research*, 47(10):834–841.

Krol, L. (2015). Distributed Monte Carlo feature selection: extracting informative features out of multidimensional problems with linear speedup. In *Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery*, pages 463–474. Springer.

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317.

Labaj, W., Papiez, A., Polanski, A., and Polanska, J. (2017). Comprehensive analysis of MILE gene expression data set advances discovery of leukaemia type and subtype biomarkers. *Interdisciplinary Sciences: Computational Life Sciences*, 9(1):24–35.

Lancaster, H. (1961). The combination of probabilities: an application of orthonormal functions. *Australian Journal of Statistics*, 3(1):20–33.

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

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161.

Lenzerini, M. (2002). Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 233–246. ACM.

Li, C.-X., Wheelock, C. E., Sköld, C. M., and Wheelock, Å. M. (2018). Integration of multi-omics datasets enables molecular classification of copd. *European Respiratory Journal*, page 1701930.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, P., Piao, Y., Shon, H. S., and Ryu, K. H. (2015). Comparing the normalization methods for the differential analysis of illumina high-throughput rna-seq data. *BMC bioinformatics*, 16(1):347.

Liptak, T. (1958). On the combination of independent tests. *Magyar Tud Akad Mat Kutato Int Kozl*, 3:171–197.

Liu, Z., Xie, M., Yao, Z., Niu, Y., Bu, Y., and Gao, C. (2013). Three meta-analyses define a set of commonly overexpressed genes from microarray datasets on astrocytomas. *Molecular neurobiology*, 47(1):325–336.

Lönnstedt, I. and Speed, T. (2002). Replicated microarray data. *Statistica sinica*, 12(1):31–46.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):550.

Luo, J., Schumacher, M., Scherer, A., Sanoudou, D., Megherbi, D., Davison, T., Shi, T., Tong, W., Shi, L., Hong, H., et al. (2010). A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *The pharmacogenomics journal*, 10(4):278–291.

Manning, G., Kabacik, S., Finnon, P., Bouffler, S., and Badie, C. (2013). High and low dose responses of transcriptional biomarkers in ex vivo X-irradiated human blood. *International journal of radiation biology*, 89(7):512–522.

Mardis, E. R. (2008). Next-generation dna sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387–402.

Masoudkabir, F., Sarrafzadegan, N., Gotay, C., Ignaszewski, A., Krahn, A. D., Davis, M. K., Franco, C., and Mani, A. (2017). Cardiovascular disease and cancer: Evidence for shared disease pathways and pharmacologic prevention. *Atherosclerosis*, 263:343–351.

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

Meehan, T. F., Vasilevsky, N. A., Mungall, C. J., Dougall, D. S., Haendel, M. A., Blake, J. A., and Diehl, A. D. (2013). Ontology based molecular signatures for immune cell types via gene expression analysis. *BMC bioinformatics*, 14(1):263.

Micallef, L. and Rodgers, P. (2014). eulerape: drawing area-proportional 3-venn diagrams using ellipses. *PloS one*, 9(7):e101717.

Mirzayans, R., Andrais, B., Scott, A., Wang, Y. W., and Murray, D. (2013). Ionizing Radiation-Induced Responses in Human Cells with Differing TP53 Status. *Int J Mol Sci*, 14(11):22409–22435.

Mistry, M. and Pavlidis, P. (2008). Gene ontology term overlap as a measure of gene functional similarity. *BMC bioinformatics*, 9(1):327.

Papiez, A., Azimzadeh, O., Tapio, S., and Polanska, J. (2019a). Integrative multiomics study for validation of mechanisms in radiation-induced ischemic heart disease. *PloS ONE*.

Papiez, A., Badie, C., and Polanska, J. (2019b). Machine learning techniques combined with dose profiles indicate radiation response biomarkers. *International Journal of Applied Mathematics and Computer Science*, 29(1).

Papiez, A., Finnon, P., Badie, C., Bouffler, S., and Polanska, J. (2014). Integrating expression data from different microarray platforms in search of biomarkers of radiosensitivity. In *IWBBIO*, pages 484–493.

Papiez, A., Kabacik, S., Badie, C., Bouffler, S., and Polanska, J. (2015). Statistical integration of p-values for enhancing discovery of radiotoxicity gene signatures. In *International Conference on Bioinformatics and Biomedical Engineering*, pages 503–513. Springer.

Papiez, A., Marczyk, M., Polanska, J., and Polanski, A. (2018). BatchI: Batch effect Identification in high-throughput screening data using a dynamic programming algorithm. *Bioinformatics*.

Park, B., Yee, C., and Lee, K.-M. (2014). The effect of radiation on the immune response to cancers. *International journal of molecular sciences*, 15(1):927–943.

Parker, H. S., Bravo, H. C., and Leek, J. T. (2014). Removing batch effects for prediction problems with frozen surrogate variable analysis. *PeerJ*, 2:e561.

Parmigiani, G., Garrett, E. S., Anbazhagan, R., and Gabrielson, E. (2002). A statistical framework for expression-based molecular classification in cancer. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):717–736.

Pietrowska, M., Polanska, J., Suwiński, R., Wideł, M., Rutkowski, T., Marczyk, M., Domińczyk, I., Ponge, L., Marczak, Ł., Polanski, A., et al. (2012). Comparison of peptide cancer signatures identified by mass spectrometry in serum of patients with head and neck, lung and colorectal cancers: association with tumor progression. *International journal of oncology*, 40(1):148–156.

Polanski, A., Marczyk, M., Pietrowska, M., Widlak, P., and Polanska, J. (2015). Signal partitioning algorithm for highly efficient gaussian mixture modeling in mass spectrometry. *PloS one*, 10(7):e0134256.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature genetics*, 32:496.

Rahimov, F., King, O. D., Leung, D. G., Bibat, G. M., Emerson, C. P., Kunkel, L. M., and Wagner, K. R. (2012). Transcriptional profiling in facioscapulohumeral muscular dystrophy to identify candidate biomarkers. *Proceedings of the National Academy of Sciences*, 109(40):16234–16239.

Ray, M., Yunis, R., Chen, X., and Rocke, D. M. (2012). Comparison of low and high dose ionising radiation using topological analysis of gene coexpression networks. *BMC genomics*, 13(1):190.

Reese, S. E., Archer, K. J., Therneau, T. M., Atkinson, E. J., Vachon, C. M., de Andrade, M., Kocher, J.-P. A., and Eckel-Passow, J. E. (2013). A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal components analysis. *Bioinformatics*, page btt480.

Reinhardt, M. J., Kubota, K., Yamada, S., Iwata, R., and Yaegashi, H. (1997). Assessment of cancer recurrence in residual tumors after fractionated radiotherapy: a comparison of fluorodeoxyglucose, L-methionine and thymidine. *The Journal of Nuclear Medicine*, 38(2):280.

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Rifai, N., Gillette, M. A., and Carr, S. A. (2006). Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nature biotechnology*, 24(8):971.

Scherer, A. (2009). *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*. Wiley Series in Probability and Statistics. Wiley & Sons, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom.

Schmid, P. R., Palmer, N. P., Kohane, I. S., and Berger, B. (2012). Making sense out of massive data by going beyond differential expression. *Proceedings of the National Academy of Sciences*, 109(15):5594–5599.

Shabalin, A. A., Tjelmeland, H., Fan, C., Perou, C. M., and Nobel, A. B. (2008). Merging two gene-expression studies via cross-platform normalization. *Bioinformatics*, 24(9):1154–1160.

Shao, L., Luo, Y., and Zhou, D. (2014). Hematopoietic stem cell injury induced by ionizing radiation. *Antioxidants & redox signaling*, 20(9):1447–1462.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Sims, A. H., Smethurst, G. J., Hey, Y., Okoniewski, M. J., Pepper, S. D., Howell, A., Miller, C. J., and Clarke, R. B. (2008). The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets–improving meta-analysis and prediction of prognosis. *BMC medical genomics*, 1(1):1.

Sîrbu, A., Ruskin, H. J., and Crane, M. (2010). Cross-platform microarray data normalisation for regulatory network inference. *PloS one*, 5(11):e13822.

Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer.

Son, C.-H., Keum, J.-H., Yang, K., Nam, J., Kim, M.-J., Kim, S.-H., Kang, C.-D., Oh, S.-O., Kim, C.-D., Park, Y.-S., et al. (2014). Synergistic enhancement of nk cell-mediated cytotoxicity by combination of histone deacetylase inhibitor and ionizing radiation. *Radiation Oncology*, 9(1):49.

Steelman, L. S., Navolanic, P., Chappell, W. H., Abrams, S. L., Wong, E. W., Martelli, A. M., Cocco, L., Stivala, F., Libra, M., Nicoletti, F., Drobot, L. B., Franklin, R. A., and McCubrey, J. A. (2011). Involvement of Akt and mTOR in chemotherapeutic- and hormonal-based drug resistance and response to radiation in breast cancer cells. *Cell Cycle*, 10(17):3003–3015.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Williams Jr, R. M. (1949). *The American soldier: Adjustment during army life.(Studies in social psychology in World War II), Vol. 1.* Princeton Univ. Press.

Sun, Z., Wu, Y., White, W. M., Donkena, K. V., Klein, C. J., Garovic, V. D., Therneau, T. M., and Kocher, J.-P. A. (2011). Batch effect correction for genome-wide methylation data with Illumina Infinium platform. *BMC Medical Genomics*, 4(1):1.

Taunk, N. K., Haffty, B. G., Kostis, J. B., and Goyal, S. (2015). Radiation-induced heart disease: pathologic abnormalities and putative mechanisms. *Frontiers in oncology*, 5:39.

Terpstra, T. J. (1952). The asymptotic normality and consistency of Kendall's test against trend, when ties arepresent in one ranking. *Proc. Kon. Ned. Akad. v. Wetensch. A*, 55:327–333.

Tini, G., Marchetti, L., Priami, C., and Scott-Boyer, M.-P. (2017). Multi-omics integration—a comparison of unsupervised clustering methodologies. *Briefings in bioinformatics*.

Trygg, J. and Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of chemometrics*, 16(3):119–128.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121.

UNSCEAR (2000). *Sources and effects of ionizing radiation.*, volume 1. United Nations Publications.

Välikangas, T., Suomi, T., and Elo, L. L. (2016). A systematic evaluation of normalization methods in quantitative label-free proteomics. *Briefings in bioinformatics*, 19(1):1–11.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001). The sequence of the human genome. *science*, 291(5507):1304–1351.

Walter, M., Bonin, M., Pullman, R. S., Valente, E., Loi, M., Gambarin, M., Raymond, D., Tinazzi, M., Kamm, C., Glöckle, N., et al. (2010). Expression profiling in peripheral blood reveals signature for penetrance in DYT1 dystonia. *Neurobiology of Disease*, 38(2):192–200.

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

Weichselbaum, R. R., Hallahan, D., Fuks, Z., and Kufe, D. (1994). Radiation induction of immediate early genes: effectors of the radiation-stress response. *International Journal of Radiation Oncology* Biology* Physics*, 30(1):229–234.

Widłak, W. (2013). High-throughput technologies in molecular biology. In *Molecular Biology*, pages 139–153. Springer.

Witek, M., Blomain, E. S., Magee, M. S., Xiang, B., Waldman, S. A., and Snook, A. E. (2014). Tumor radiation therapy creates therapeutic vaccine responses to the colorectal cancer antigen gucy2c. *International Journal of Radiation Oncology* Biology* Physics*, 88(5):1188–1195.

Xi, H., Shulha, H. P., Lin, J. M., Vales, T. R., Fu, Y., Bodine, D. M., McKay, R. D., Chenoweth, J. G., Tesar, P. J., Furey, T. S., et al. (2007). Identification and characterization of cell type–specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS genetics*, 3(8):e136.

Xu, X., Wells, A. B., O'Brien, D. R., Nehorai, A., and Dougherty, J. D. (2014). Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders. *Journal of Neuroscience*, 34(4):1420–1431.

Yarnold, J., Ashton, A., Bliss, J., Homewood, J., Harper, C., Hanson, J., Haviland, J., Bentzen, S., and Owen, R. (2005). Fractionation sensitivity and dose response of late adverse effects in the breast after radiotherapy for early breast cancer: long-term results of a randomised trial. *Radiotherapy and oncology*, 75(1):9–17.

Yi, H., Raman, A. T., Zhang, H., Allen, G. I., and Liu, Z. (2017). Detecting hidden batch factors through data-adaptive adjustment for biological effects. *Bioinformatics*, 34(7):1141–1147.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

Zaykin, D. V. (2011). Optimally weighted z-test is a powerful method for combining probabilities in meta-analysis. *Journal of evolutionary biology*, 24(8):1836–1841.

Zhan, Q. (2005). GADD45A, a p53-and BRCA1-regulated stress protein, in cellular response to DNA damage. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 569(1):133–143.

Zhu, B., Song, N., Shen, R., Arora, A., Machiela, M. J., Song, L., Landi, M. T., Ghosh, D., Chatterjee, N., Baladandayuthapani, V., et al. (2017). Integrating clinical and multiple omics data for prognostic assessment across human cancers. *Scientific reports*, 7(1):16954.

Zomer, A., Vendrig, T., Hopmans, E. S., van Eijndhoven, M., Middeldorp, J. M., and Pegtel, D. M. (2010). Exosomes: fit to deliver small rna. *Communicative & integrative biology*, 3(5):447–450.

List of Author's Publications

JCR Indexed Articles

Papiez, A., Badie, C., and Polanska, J. (2019). Machine learning techniques combined with dose profiles indicate radiation response biomarkers. *International Journal of Applied Mathematics and Computer Science*, 29(1) **(MNiSW: 25 points, IF: 1.649)**

Papiez, A., Azimzadeh, O., Azizova, T., Moseeva, M., Anastasov, N., Smida, J., Tapio, S., and Polanska, J. (2018a). Integrative multiomics study for validation of mechanisms in radiation-induced ischemic heart disease in mayak workers. *PLOS ONE*, 13(12):1–14 **(MNiSW: 35 points, IF: 2.766)**

Papiez, A., Marczyk, M., Polanska, J., and Polanski, A. (2018b). BatchI: Batch effect Identification in high-throughput screening data using a dynamic programming algorithm. *Bioinformatics* **(MNiSW: 45 points, IF: 5.481)**

Labaj, W., Papiez, A., Polanski, A., and Polanska, J. (2017). Comprehensive analysis of MILE gene expression data set advances discovery of leukaemia type and subtype biomarkers. *Interdisciplinary Sciences: Computational Life Sciences*, 9(1):24–35 **(MNiSW: 15 points, IF: 0.796)**

Web of Science Indexed Conference Articles

Labaj, W., Papiez, A., Polanska, J., and Polanski, A. (2016). Deep data analysis of a large microarray collection for leukemia biomarker identification. In *10th International Conference on Practical Applications of Computational Biology & Bioinformatics*, pages 71–79. Springer **(MNiSW: 15 points)**

Papiez, A., Kabacik, S., Badie, C., Bouffler, S., and Polanska, J. (2015). Statistical integration of p-values for enhancing discovery of radiotoxicity gene signatures. In *International Conference on Bioinformatics and Biomedical Engineering*, pages 503–513. Springer **(MNiSW: 15 points)**

Papiez, A., Finnon, P., Badie, C., Bouffler, S., and Polanska, J. (2014). Integrating expression data from different microarray platforms in search of biomarkers of radiosensitivity. In *IWBBIO*, pages 484–493 **(MNiSW: 10 points)**

Monograph Chapters

Papież, A., Skrzypski, M., Szymanowska-Narloch, A., Jassem, E., Maciejewska, A., Pawłowski, R., Dziadziuszko, R., Jassem, J., Rzyman, W., and Polańska, J. (2018). Can an integrative SNP approach substitute standard identification in comprehensive case/control analyses. In Fdez-Riverola, F., Mohamed, M., Rocha, M., De, P., and Gonzalez, P., editors, *Advances in Intelligent Systems and Computing*, volume 803, pages 123–130. Springer, Cham **(MNiSW: 5 points)**

Papież, A., Badie, C., and Polańska, J. (2017b). Response profiles for high and therapeutic radiation doses in breast cancer patients. In Forys, U. and Śmieja, J., editors, *Proceedings of the XXIII National Conference on Applications of Mathematics in Biology and Medicine*, pages 137–142. Silesian University of Technology, Gliwice, Poland **(MNiSW: 5 points)**

Conference Proceedings Abstracts

Papież, A., Azimzadeh, O., Tapio, S., and Polańska, J. (2017a). Regression analysis for dose and age related deregulated proteins. In *4th ICRP Symposium on the system of radiological protection and 2nd European Radiological Protection Research Week*, page 83

Papież, A., Żyła, J., Binczyk, F., and Polańska, J. (2017d). Drosophila melanogaster RNA-Seq analysis by k-means clustering with adaptive initial conditions. In *XXI Gliwice Scientific Meetings*, page 134

Papież, A., Danek, A., Gruca, A., Łabaj, P., and Polańska, J. (2017c). Functional annotation differences among Drosophila melanogaster strains. In *Computational Approached in Precision Medicine*, page 30

Łabaj, W., Papież, A., Polańska, J., and Polański, A. (2017). Obszerna analiza danych wysoce zrównoleglonych dla identyfikacji biomarkerów białaczki. In *IV Śląskie Spotkania Naukowe*, page 28

Łabaj, W., Papież, A., and Polańska, J. (2016). Leukemia subtype biomarker validation in a large gene expression study. In *XXth Gliwice Scientific Meetings*

Papież, A., Badie, C., and Polańska, J. (2016a). An integrative approach vs restrictive thresholds for combining gene expression data sets on radiation response. In *2nd Congress of Polish Biochemistry, Cell biology, Biotechnology and Bioinformatics*, page 170

Papież, A., Marczyk, M., Polański, A., and Polańska, J. (2016b). Dynamic programming as a batch effect identification method (Programowanie dynamiczne jako metoda identyfikacji efektu paczki). In *III Śląskie Spotkania Naukowe*, page 35

Tobiasz, J. and Papież, A. (2015). The application of the microarray analysis methods in search of candidate gene signatures of radiosensitivity. In *XIXth Gliwice Scientific Meetings*, page 151

Papież, A., Badie, C., and Polańska, J. (2015b). P-Value Integration as a technique for validating high-throughput biomedical experiments. In *International Synthetic and Systems Biology Summer School*, page 32

Papież, A., Badie, C., and Polańska, J. (2015a). Merging high-dimensional data at the p-value level as a superior solution to entire data set fusion. In *8th Symposium of the Polish Bioinformatics Society*, page 122

Papież, A., Marczyk, M., Polański, A., and Polańska, J. (2015). Identifying batch effects in high-throughput biological data using dynamic programming based approach. In *19th Annual International Conference on Research in Computational Biology*

Papież, A., Badie, C., and Polańska, J. (2014b). Statistical methods for integrating high-throughput biological data. *ACTA BIOCHIMICA POLONICA*, page 95

Papież, A., Badie, C., and Polańska, J. (2014a). Impact of the selection of statistical test on the quality of gene signatures in an integrative analysis approach. In *XVIIIth Gliwice Scientific Meetings*

Papież, A. (2014a). Integrative bioinformatics in radiosensitivity analysis (Bioinformatyka integratywna w analizie radiowrażliwości). In *III Śląskie Spotkania Naukowe*

Papież, A. (2014b). Statistical data integration as a tool for combining information derived from biological experiments (Statystyczna integracja danych jako narzędzie łączenia informacji pozyskanych w eksperymentach biologicznych). In *I Seminarium Polskiego Towarzystwa Proteomicznego*

Papież, A., Finnon, P., Bouffler, S., Badie, C., and Polańska, J. (2013). Methods for meta-analysis of expression data from different microarray platforms. In *XVIIth Gliwice Scientific Meetings*, page 181