Natascha Fladischer BSc

To integrate, or not to integrate . . .

A comparison between sequential and integrative analysis methods for omics data

MASTER'S THESIS

to achieve the university degree of

Diplom-Ingenieurin

Master's degree programme: Biomedical Engineering

submitted to

Graz University of Technology

Supervisor

Dr. Gerhard Thallinger

Institute of Computational Biotechnology

8010 Graz, Petersgasse 14/V, Austria

Institute of Neural Engineering

8010 Graz, Stremayrgasse 16/IV, Austria

Graz, 09 November 2017

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Date Signature

Acknowledgement

I am very thankful to my supervisor Dr. Gerhard Thallinger, who gave me the opportunity to do my master thesis on this topic, as well as to be the examiner of this master thesis. I also thank him very much for his expertise and advice, which enabled me to complete my master's thesis successfully, and also for the critical reading of the comprehensive work. Without his support this work would not have been possible.

Furthermore, I would like to thank Dr. Oana Alina Zeleznik for her valuable and accompanying support throughout this master thesis. She made a considerable contribution to the conceptual design of this master thesis.

Last but not least, I would like to thank my family, who has always supported me during the years of my studying.

Contents

Abstract

System-wide molecular profiling studies measure biological samples and organisms on a global scale, on various types of biological molecules, for example on DNA, RNA, protein and metabolite level. This results in high-dimensional omics data (genomics, transcriptomics, proteomics and metabolomics) obtained from high-throughput technologies. Dimension reduction techniques, such as principal component analysis (PCA), correspondence analysis (CA) and non-symmetric correspondence analysis (NSCA), are very important tools for the analysis of this data but they are limited to the analysis of a single level. Therefore, integrative analysis methods, such as co-inertia analysis (CIA) and multiple co-inertia analysis (MCIA), have been developed which allow simultaneous analysis of two or more data sets. MCIA is a powerful tool in the analysis of multiple high-dimensional data sets because this method enables the visualization of measurements/samples using a lower number of features/variables and therefore, facilitates the detection, representation and biological interpretation of the correlated structure within and between the different data sets. This study focuses on the comparison of sequential (PCA, CA, NSCA) and integrative analysis (MCIA) methods based on data comprising 60 human cancer cell lines at the transcript and protein level.

Surprisingly, PCA and NSCA do not show considerably higher degree of divergence among gene and protein expression data. These two methods demonstrate mainly inconsistent clustering based on protein expression. It is noteable that CA based on proteome level provided a more homogenous clustering compared to PCA and NSCA. MCIA proves that protein and gene expression profiles can be regarded as powerful molecular descriptors of different cancer tissues and integrative analysis of both provides a deeper insight on multiple layers of biological systems compared to any analysis tool alone. The most homogenous clustering and separation of the different carcinomas according to the most influential genes was achieved with the MCIA, and the different types of cancer were thus clear separated on the first two principal components. The most influential genes based on MCIA at each end of the first principal component and the second principal component provided useful information regarding the clustering of the individual cancer types. Thus, MCIA provide more information and is more powerful than corresponding sequential. Furthermore, MCIA achieves a homogenous separation of mesenchymal (glioblastoma and renal cancer cells) and epithelial (leukemia and colon cancer cells) markers which promote epithelial-mesenchymal transition (EMT) that has a significant function in cancer biology and is also involved the malignancy and metastasis of epithelial cancer cells. Mesenchymal cells have migratory and invasive characteristics which have a significant function in malignant metastasizing cancer. Thus, MCIA also yields important information regarding invasive, metastatic phenotype in cancer cells.

Kurzfassung

Die systematische Analyse von biologischen Organismen soll dazu beitragen biologische Prozesse in ihrer Gesamtheit zu verstehen. Das Ziel ist, ein integriertes Bild aller Prozesse auf unterschiedlichen Ebenen von Biomolekülen, beispielsweise auf DNA-, RNA-, Protein- und Metabolit-Ebene, zu erhalten. Diese Art der Analysen werden mittels Hochdurchsatztechnologien durchgeführt und führen zu hochdimensionalen omics-Daten (Genomik, Transkriptomik, Proteomik und Metabolomik). Zur sinnvollen Analyse dieser mehrdimensionalen Daten kommen Dimensionsreduktionstechniken, wie Hauptkomponentenanalyse (HKA), Korrespondenzanalyse (CA) und nicht symmetrische Korrespondenzanalyse (NSCA), zum Einsatz. Wobei die Anwendung dieser Techniken nur auf einzelne Ebenen limitiert ist. Daher wurden integrative Analysemethoden wie Co-Inertia-Analyse (CIA) und Multiple Co-Inertia-Analyse (MCIA) entwickelt, die eine gleichzeitige Analyse von multiplen Datensätzen erlauben. MCIA ist ein leistungsfähiges Werkzeug bei der Analyse mehrerer hochdimensionaler Datensätze, da diese Methode die Visualisierung von Messungen/Proben mit einer geringeren Anzahl von Merkmalen/Variablen ermöglicht und somit die Erkennung, Darstellung und biologische Interpretation der korrelierten Struktur innerhalb und zwischen den verschiedenen Datensätzen erleichtert. Diese Arbeit fokussiert sich auf den Vergleich von sequentiellen (HKA, CA, NSCA) und integrativen Analyse (MCIA) Methoden. Als Model dienten Transkript- und Protein-Daten von 60 menschlichen Krebszelllinien.

Überraschenderweise zeigten die HKA und die NSCA keine signifikant höhere Divergenz zwischen Gen- und Protein-Expressionsdaten. In beiden Fällen lieferte das Clustering basierend auf den Protein-Expressionsdaten inkonsistente Ergebnisse. Im Unterschied dazu lieferte das Clustering der Proteom-Daten im Falle der CA basierend auf Protein-Expressionsdaten vergleichbarere Ergebnisse. Die MCIA zeigt, dass Protein- und Genexpressionsprofile als geeignete molekulare Deskriptoren verschiedener Krebsgewebe angesehen werden können und eine integrative Analyse von beiden einen tieferen Einblick in die Gesamtheit eines biologischer Systeme im Vergleich zu einzelnen Analysewerkzeugen bietet. Mittels MCIA konnte eine homogene Clusterbildung und Trennung der verschiedenen Krebstypen nach den einflussreichsten Genen erreicht werden. Dadurch wurden die verschiedenen Krebsarten klar auf den ersten beiden Hauptkomponenten getrennt. Die einflussreichsten Gene, die auf MCIA an jedem Ende der ersten Hauptkomponente und der zweiten Hauptkomponente basieren, lieferten nützliche Informationen über die Clusterbildung der einzelnen Krebsarten. So bietet MCIA mehr Informationen und ist leistungsfähiger als entsprechende sequentielle Analysenmethoden. Darüber hinaus erreicht MCIA eine homogene Trennung von mesenchymalen (Glioblastom- und Nierenkrebs) und epithelialen (Leukämie- und Darmkrebs) Markern, die den epithelial-mesenchymalen Übergang fördern. Dieser hat nachweislich eine signifikante Funktion in der

Krebsbiologie, als auch der Malignität und Metastasierung von Epithelkrebszellen. Mesenchymale Zellen haben wandernde und invasive Eigenschaften, die eine signifikante Funktion bei malignen metastasierenden Krebstypen haben. Somit lassen sich mittels MCIA auch wichtige Informationen über den invasiven, metastatischen Phänotyp der unterschiedlichen Krebszellen eruieren.

1 Introduction

1.1 The world of omics technologies

The fast progression in life science research presents big challenges in the development of novel technologies to answer complex biological questions. Tremendous progress in the development of novel analytic methods over the past decades enabled the generation of huge complex biological data sets. These new high-throughput analytical technologies facilitate the simultaneous measurement of a large number of molecules, as well as possible alterations. The analysis of biological samples can occur on DNA, RNA, protein or metabolite level. In the end, the generated data should allow to get novel and systematic insights into biological systems to answer relevant questions. Typical examples include the comparative measurement of gene expression levels with microarrays and RNA sequencing (RNA-seq) or the comparative measurement of protein expression levels. The term -omics is often used in research fields that denote studies undertaken on a large scale. Depending on the molecules investigated different types are known, such as genomics, transcriptomics, proteomics and metabolomics. The Human Genome Project forms the basis for the development of novel techniques in this field and resulting in the generation of a tremendous amount of novel data [1]. This is also demonstrated by the omics revolution during the early 21^{th} century. Due to these developments it is nowadays possible to generate a large number of measurements over a short time period providing data about the DNA of organisms (genomics), transcribed RNA of genes over time (transcriptomics), protein profiles of particular tissues or cells (proteomics) and metabolites (metabolomics) to a broad scientific community [2].

The central dogma of molecular biology comprises a defined flow of information [3]. Starting at the transcription of a gene from DNA to mRNA, over the translation of the protein, to different metabolites produced by different enzymatic activities, each individual step can be tackled by one of the available omics technologies (see Figure 1). Within the next section a short overview of the different approaches will be given.

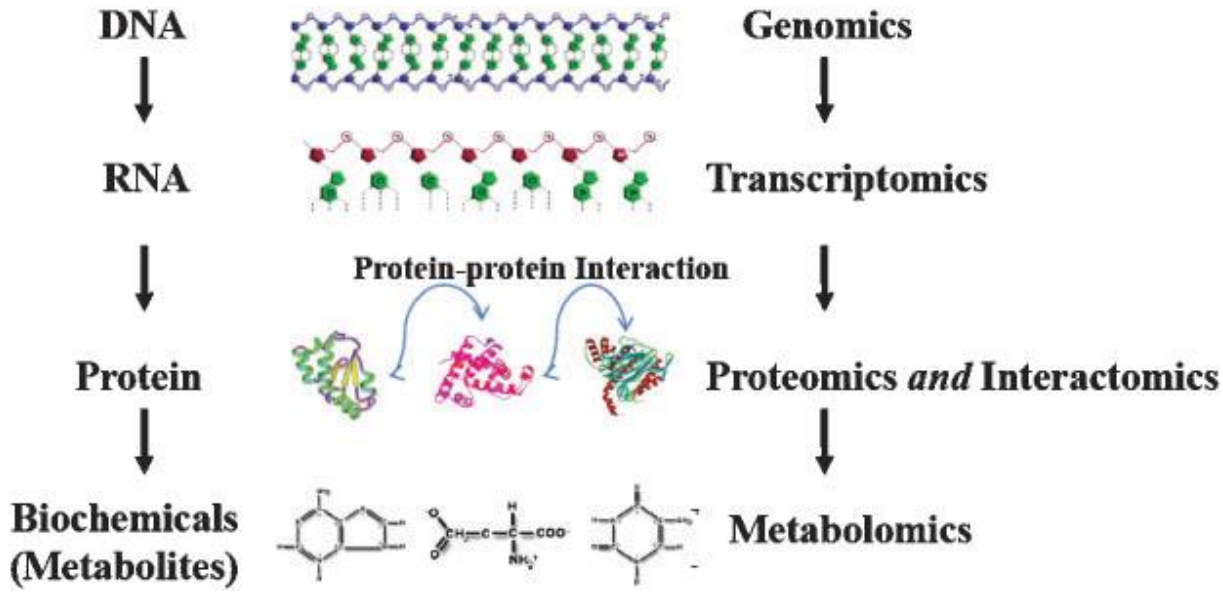


Figure 1: Schematic overview of the information flow within a cell and the corresponding omics technology for the analysis of each individual step [3].

1.1.1 Genomics

Genomics is a discipline that studies the complete genetic material (genome) of an organism, as well as its function [4]. Since this discipline investigates the entirety of all genes and the interactions between them, it has the ability to identify the common impact of the genetic material on basic mechanisms, such as growth and development, of a biological system [4]. The main goal of genome-wide analysis is the identification of DNA sequences, mapping of the genome structure, as well as functional prediction of genes. The resulting genetic information is collected and stored in databases, like NCBI [5] and ENSEMBL [6] and therefore accessible for the public domain.

Genomics was the primary field investigated in the era of -omics for which high-throughput analytical measurements were obtainable. Therefore, various methods have been developed and applied in genomics, such as DNA microarrays for comparative genomic hybridization (CGH) or determination of single-nucleotide polymorphism (SNP) [7]. These DNA microarrays (also called gene chips) facilitate also the measurement of RNA-expression levels (see 1.1.3 Transcriptomics). Microarray-based CGH is used to investigate genomic alterations in different diseases [8]. These CGH arrays enable the detection of copy number variations [9], including genomic deletions, genomic duplications and complex genomic rearrangements [10]. The schematic representation of a CGH-array technology is shown in [8]. An alternative application is represented by SNP arrays [7]. This method enables the simultaneous genotyping of DNA based on single nucleotide exchanges (SNPs) [11,12]. These high-throughput technologies

provide a better understanding of a mechanism of various human diseases on the molecular level, since a considerable amount of copy number variants (CVNs) are associated with complex diseases [9]. Therefore CHG and SNP technologies have expanded their application to various fields of research, such as medicine, in particular cancer research [13,14]. Other high-throughput technologies, which belong to the group of sequencing techniques, are summarized in [14]. This review provides an overview of the development from first generation DNA sequencing to the new developed high-throughput next-generation sequencing (HT-NGS) technologies.

Although genomic studies helped to solve a range of biological questions and get a better understanding of molecular processes, the method is limited to static DNA sequences. Therefore dynamic processes, like the metabolism and physiological state of an organism, cannot be tackled. These limitations aim for further improved methods for measuring different cellular molecules, e.g. mRNA, proteins and metabolites.

1.1.2 Epigenomics

Coming from the static DNA world, the next goal in the post-genomic era was to study dynamic processes. The omics field based on this research area was called epigenomics. This field studies alterations in the genome excluding DNA sequence mutations [15]. It deals with the entirety of the epigenetic modifications on the genetic material of a cell, also referred to as epigenome.

The two most common epigenetic modifications are DNA methylation [16], the covalent modification of cytosine, and histone modifications [17], such as methylation, acetylation and phosphorylation [18].

Epigenetic modifications influence a variety of cellular processes including cell differentiation and development, as well as the regulation of gene expression, which result in profound epigenetic alterations through various cell types [19–21]. Compared to the genome, which is relatively static in a biological system, the epigenome is more dynamic because it can be influenced by various environmental conditions, as well as lifestyle factors [22]. Furthermore a variety of studies show an association between epigenomic data and complex diseases, such as cancer [23,24].

One of the earlier and commonly used technologies for the detection of DNA methylation is represented by bisulfite genomic sequencing which was first introduced by Frommer *et al.* [25]. This standard application is used because it represents a qualitative, quantitative and efficient method for the determination of 5-methylcytosine at single base-pair resolution [26]. Several

other technologies have been developed with respect to the working basis of bisulfite, such as Methylation Specific PCR (MSP) [27], Combined Bisulfite Restriction Analysis (COBRA) [28] and Methylation-sensitive Single Nucleotide Primer Extension (Ms-SNuPE) [29].

Several different large-scale technologies have been developed for the analysis of histone modifications. These technologies are mainly based on chromatin immunoprecipitation, also known as ChIP [30,31]. This tool is often used to investigate the relationship of specific transcription factors and modified histones with the endogenous DNA region [30]. In addition it will be used for the investigation of protein-DNA interactions at a genome-wide scale [30]. More details regarding ChIP can be found in [30]. ChIP has been extended in order to obtain two other tools called ChIP-chip [32], which means a coupling of ChIP with DNA microarray (chip), and ChIP sequencing technology [33], also known as ChIP-seq, which signifies a coupling of ChIP with massively parallel DNA sequencing. The ChIP-chip method enables to detect the whole spectrum of DNA binding sites for a protein of interest in living cells using hybridization of the enriched fraction on a microarray [32]. The ChIP-seq technology is also commonly used for *in vivo* mapping the genomic location of transcription-factor binding and histone modifications [33].

1.1.3 Transcriptomics

The next molecule tackled was the RNA, the main target in a new developed research field called transcriptomics. Transcriptomics is the quantitative study of the transcriptome, which refers to the entirety of RNA transcripts in a cell or organism [34,35]. These known RNA transcripts include messenger ribonucleic acids (mRNAs) and non-coding ribonucleic acid (ncRNAs) [36], such as transfer ribonucleic acid (tRNA) and ribosomal ribonucleic acid (rRNA), transcribed from the genome of an organism [35]. The three mentioned RNA molecules (mRNA, tRNA and rRNA) play an essential role in protein synthesis [37]. Another group among the non-coding RNAs is called microRNAs (miRNAs). Further details regarding miRNAs can be found in [38].

Transcriptomic research aims to catalog all types of RNA transcripts, such as mRNAs and non-coding RNAs. In addition, transcriptomics deals with the identification of the transcriptional structure of genes regarding their transcription start sites, mRNA splicing patterns and other post-transcriptional modifications. Further on, it is of great interest to quantify changes in gene expression levels of all RNA transcripts under various environmental conditions and during various stages of development (reviewed by Wang *et al.* [39]).

Transcriptomics provide insight in the gene activity (gene expression) at the transcriptional level in order to obtain a deeper understanding of different molecular mechanisms, such as gene structure, regulation of gene expression, as well as gene product function [40]. Furthermore several studies show a direct correlation between transcriptomic data and various diseases, particularly in cancer [41].

Several methods have been developed in order to investigate the transcriptome. Hybridization- or sequence-based technologies are used to analyze the transcriptome of organisms [39]. Hybridization-based technologies include microarrays provided by Affymetrix [42,43], spotted complementary deoxyribonucleic acid (cDNA) [44] or oligonucleotide array technology [45]. A comparison between these three microarray technologies is described in [46].

A typical cDNA microarray experiment is described in detail by Liang and coworkers and it is performed in the following manner [47]: firstly, an abundance of cDNA clones (probes) are amplified by PCR and spotted on a glass slide with the aid of a robotic arrayer. Next, the RNA of two samples (patient/control or test/reference) is reverse transcribed in order to generate the corresponding cDNA. The cDNA is then labeled with two fluorescent dyes, such as cyanine dyes, e.g. Cy3 and Cy5, which are representing different samples. These two fluorescent dyes are widely used and combined in two-color microarray experiments. After that the labeled cDNA is hybridized to the microarray. Finally, the unbound cDNA is washed from the glass slide. A laser confocal scanner measures the fluorescence intensities in each spot. These intensity data are analyzed in order to obtain the mRNA expression profiles (see Figure 2). A microarray experiment depends on several further steps, such as image analysis, data preprocessing, normalization, data analysis, standardization, data storage, data integration and biological interpretation.

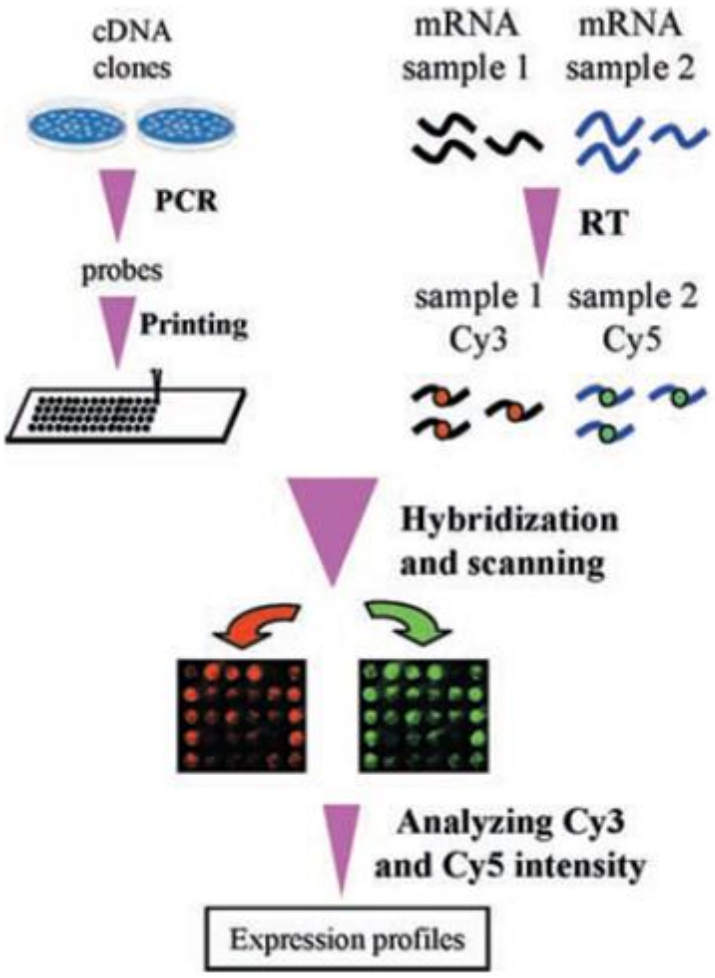


Figure 2: Basic procedure and schematic representation of a typical cDNA microarray experiments (Figure originating from [47]).

In contrast, sequence-based technologies enable the direct sequencing of the cDNA. One of the earlier techniques was Sanger sequencing [48] of expressed sequence tag (EST) libraries [49]. On the one hand the Sanger technology offers high accuracy and long read length, but on the other hand it suffers from high costs, low-throughput and non-quantitative results [39,50]. Further improvements led to the development of tag-based technologies [51] including serial analysis of gene expression (SAGE) [52], cap analysis of gene expression (CAGE) [53] and massively parallel signature sequencing (MPSS) [54]. These methods enable the quantification of gene expression levels of thousands of genes simultaneously and therefore they are also called high-throughput technologies. The latest developed method is the RNA-seq technology [39,55].

A typical RNA-seq process is performed in the following manner [39]: RNA transcripts are reverse transcribed into a library of cDNA fragments. Adaptors are ligated to any cDNA fragment. A variety of high-throughput technologies, such as Illumina, Roche and Applied Biosystems Solid (all three reviewed by [14]), can be used for RNA-seq in order to sequence each molecule (amplified or not amplified) in high-throughput. As a result short sequences, reads ranging from 30 to 400 base pairs (bp) depending on the used high-throughput method, are

generated through single-end sequencing or pair-end sequencing. RNA-seq has some advantages over other high-throughput technologies (see Table 1) and therefore it is one of the most currently and used technologies for transcriptome analysis [39]. Further details regarding RNA-seq can be found in [39].

Table 1: Advantages of RNA-seq compared to other high-throughput technologies which are also used for transcriptome analysis (Table originating from [39])

Technology	**Microarray**	**cDNA or EST sequencing**	**RNA-seq**
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 base pairs (bp)	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited to gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundred fold	Not practical	> 8 000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
cost for mapping transcriptomes of large genomes	High	High	Relatively low

1.1.4 Proteomics & Interactomics

Proteins, which are translated from coding transcripts, are primary components of a living organism and play an important role in physiological states and metabolic pathways of cells. To get a better understanding of structure and function of proteins within a single cell or an organism, the field of proteomics studies the entirety of expressed proteins (proteome) [34]. Furthermore, a more detailed characterization of protein expression, localization, biochemical activity, abundance, interactions, modifications and cellular roles of proteins in their natural environment is possible [56]. This data helps to understand cellular processes.

Initially, proteomics was used for the identification of the whole set of proteins expressed by a cell, tissue or an organism [57]. Other goals of the current research focus are the determination of different properties of proteins, such as quantity, modification and structure are in the current research focus [57]. Additionally, proteomics provides a deeper understanding of protein expression and its alterations under the impact of biological disorders, such as diseases or drug

treatment. Moreover, researcher focused on the detection of alterations in the proteome regarding changes in cellular state, such as differentiation or different environmental conditions.

Nowadays there are three different types of proteomic approaches in use: protein expression proteomics, structural proteomics and functional proteomics [58]. In protein expression proteomics global alterations in protein expression levels between two or more samples are measured. This discipline is also known as expression proteomics. Functional proteomics is a very diverse field focusing on various issues, such as the isolation of protein complexes. This approach includes the isolation of specific protein types by the use of protein ligands or affinity chromatography. This helps to get a better understanding of the role of proteins in signal transduction or disease mechanisms.

Proteins are dynamic molecules in the cell which means that they can vary in one host over time, but as well host to host alterations are possible. These alterations can include attachment to the cell membrane, protein synthesis and degradation. In addition, proteins are subject to one or more post-translational modifications, including phosphorylation, acetylation, methylation, glycosylation, ubiquitination, nitrosylation [59].

Several high-throughput technologies to investigate protein expression and modification have been developed in order to measure proteins on a large scale. These approaches enable simultaneous analysis of the expression level of proteins within a cell, system or organism. Such tools are 2D gels which have the ability to separate and visualize proteins by two physical properties, e.g. charge and mass [60,61]. Mass spectrometry (MS) allows the identification of protein sequence [62,63] and the quantification of the protein [64]. Nuclear magnetic resonance (NMR) spectroscopy is another efficient technology for the determination of the protein structures [65]. ChIP-seq can be applied to study *in vivo* protein-DNA interactions [66]. Fluorescence imaging enables the localization of a protein within a cell [67].

An extension of proteomics is interactomics. It deals with the identification of protein-protein interactions (PPIs), the investigation of their individual components and the impact of these interactions between proteins, as well as other cell components [68]. These PPIs are of central importance for the understanding of different cellular functions.

Various high-throughput experimental approaches have been developed, such as Yeast-2-hybrid systems [69,70] and high-throughput chips [71].

1.1.5 Metabolomics

Metabolomics is the global study of small-molecule metabolite profiles within a cell, tissue or organism under defined conditions [34]. The metabolome is made up by the total amount of low molecular weight molecules (metabolites) involved in metabolic reactions within a biological system (cell, tissue or organism) [34]. The metabolome is built up of molecules from very diverse and different groups and therefore it is the most complex (in physical and chemical terms) class to be studied by omics technologies [34]. Metabolites, such as lipids, alcohols, amino acids, nucleotides, antioxidants, organic acids and vitamins [72], participate in various intrinsic chemical processes within cells and are influenced by different environmental conditions.

Two different state of the art applications to study the metabolome of an organism are used: targeted and untargeted approaches [73]. Firstly, the targeted analysis only tackles a small number of known metabolites (targets) using a suitable analytical method specific for the detection of the metabolites of interest [74]. In contrast, untargeted metabolomic analysis intends to determine the entire metabolome present in a biological system, as well as useful information of unknown compounds which make up the majority of all detected compounds [73]. Further details regarding the targeted and untargeted approaches for metabolome analysis can be found in [73].

Driven by the technological advances made in different omics field, significant improvements in measuring the whole set of metabolites have been made. The separation and detection of metabolites rely on state-of-the-art chromatography devices coupled with different detectors. Separation can be performed by gas chromatography (GC) [75], liquid chromatography (LC) [76], high-performance liquid chromatography (HPLC) [77], ultra-performance liquid chromatography (UPLC) [78] and capillary electrophoresis (CE) [79]. In many cases this separation tools are coupled to a MS for detection [80]. Alternatively, NMR spectroscopy can be used [81].

1.1.6 Cross-omics fields

The single omics analysis can very often not fully unravel the complexity of a biological system (cell, tissue or organism). That is the reason why it is meaningful to consider organisms as whole complex biological systems and therefore analyze multi-level omics data because each omics technology reflects a single level of the biological state of the cell.

The data of high-throughput technologies are widely used in various research fields, such as biomedicine and for their analysis different bioinformatic tools are needed. In a variety of research areas multi-omics data are used, for example, for getting a deeper understanding of human diseases [82]. Some bioinformatics tools have evolved which facilitate the integration of different omics types, such as integrOmics [83] and omicade4 [84–86].

1.2 Analysis methods for omics data

In the pre-omics era the extent of data was small enough to analyze it manually. The recent developments in omics technologies allow the generation of huge complex data sets in a short period of time. So the bottleneck in laboratories refers no longer to data generation, but rather the data management, data analysis, data integration and data interpretation [87]. Therefore scientists are focusing on developing various multivariate computational analysis methods which integrate a large quantity of data sets with various omics types. These methods should tackle the high-dimensionality of the data sets. Therefore, they can be applied to data with small sample size (measurements) and a comparatively large number of variables (features), which is characteristically for omics studies.

Due to the availability of these data sets it was necessary to develop new suitable analysis methods. These methods aim to integrate the multiple data sets in a meaningful way and are called integrative analysis methods [88,248]. The methods are based on a single analysis framework because the analyses of multiple omics data sets are performed simultaneously.

The available omics data sets determine which of the two scenarios regarding multiple omics data analysis can be performed. In the first case the data sets can be measured on various omics levels but stemming from the same biological sample [89]. In the second case they can be measured on the same omics level but stemming from various organisms [90].

Integrative analysis tools provide a deeper insight and comprehensive understanding of a biological system throughout the integration of multi-level omics data. A few examples are

mentioned below. A study in which these methods have been applied to transcriptomics and proteomics data of *Plasmodium falciparum* life cycle was performed by Tomescu *et al.* [91]. There are also many other studies which analyze multiple omics data by integrating transcriptomics and proteomics data [92], such as the transcript and protein levels across the *Plasmodium falciparum* life cycle [93]. Furthermore Wang *et al.* [94] have reviewed recent advances in the integrative analysis of gene expression and methylation data. Another paper focuses on the integrative analysis of genotypic variations at the DNA level and gene expression alterations at the RNA level related to mouse weight [95]. Other studies deal with the integrative analysis of somatic mutations, copy number and gene expression levels in integrated cancer genomic data [96], chromatin maps and gene expression profiles in *Drosophila* [97] or ChIP-seq and RNA-seq data of the human regulatory network derived from ENCODE data [98]. Tebaldi *et al.* demonstrated the useful combination of mRNA, microRNA and protein expression data in mammalian cells [99].

Dimension reduction techniques (DRTs) such as principal component analysis (PCA) [88,100–103], correspondence analysis (CA) [104–107] and non-symmetric correspondence analysis (NSCA) [106,108,109] were already used in earlier exploratory data analysis and were applied to analyze omics studies involving transcriptomics and proteomics data [89,101–103,105]. These DTRs enable the projection of data sets, consisting of measurements (samples, arrays) and features (genes, mRNAs, microRNAs, proteins, metabolites), to a lower dimensional space. The resulting visualization of the data sets facilitates the detection of the most variant features through the available data sets and simplifies the interpretation of global variance between data sets. A disadvantage of DRTs is that they cannot be applied to multiple data sets simultaneously.

Therefore improved statistical methods have been developed which facilitate the integration of two data sets and enable the simultaneous analysis of two data sets in parallel [108]. The main challenge is still to extract the important biological information from multiple omics data. A method that enables the coupling of two data sets from a data matrix and tries to extract the discernible biological meaning from two omics data sets, is called co-inertia analysis (CIA) [88]. Initially, CIA was applied to ecological data tables [110], such as shown by Doldec *et. al* [111]. In their study they investigated the relationship between species and their environment. With time CIA was also applied to other data, such as genomic data, as investigated by Culhane *et al.* in a comparative study of two microarray experiments, Affymetrix and spotted cDNA array technology [112]. Recently, CIA was applied to proteomic and gene expression data where an integrative analysis on two publicly available data sets, the *Plasmodium falciparum* life cycle

data set [91] and the NCI-60 data set [89], was performed. Furthermore, there are many other integrative analysis methods, which integrate two data sets, such as canonical correlation analysis (CCA) [113,114], canonical correspondence analysis (CCA) [115,116] and generalized singular value decomposition (gSVD) [90,117]. All the above-mentioned techniques are limited in their application to only two data sets and therefore they have a limited use in modern high-dimensional multiple omics analysis.

To overcome this limitation, several different integrative analysis methods have been developed which facilitate the integration of more than two data sets and enable the simultaneous analysis of multiple omics data sets, including regularized generalized canonical correlation analysis (RGCCA) [118], sparse generalized canonical correlation analysis (SGCCA) [119], penalized canonical correlation analysis [120] and consensus PCA [121]. Another method of integrative analysis is based on multiple co-inertia analysis (MCIA) [86,88].

MCIA can simultaneously project data sets onto the same low dimensional space and therefore identify the most variant features by maximizing the square covariance among eigenvectors [122]. MCIA can detect significant features even when feature annotations are not available through the entire data set [86]. To the best of our knowledge no direct comparison between sequential and integrative analysis methods has been performed yet. The hypothesis behind the use of these methods is that integrative analysis is more powerful than corresponding sequential analysis because integrative analysis methods allow a deeper insight and comprehensive understanding of a biological system due to the integration of multi-level omics data, which means multiple layers of information. In contrast, corresponding sequential analysis methods focus on the analysis of single data sets.

1.3 Aims of the study

The overall goal of this thesis is the comparison of the results of MCIA (an integrative method) with those of three sequential methods: PCA, CA and NSCA applied to a data set comprising 60 human cancer cell lines (NCI-60) on the transcriptome and proteome level.

Specifically the following should be achieved:

- Literature research on sequential and integrative analysis methods for omics data.
- Characterization of the NCI-60 data set and re-analysis and re-annotation of the gene expression data to select the most appropriate microarray platform for the subsequent analysis.
- Data sets of gene and protein expression from the NCI-60 data should be investigated using traditional approaches, such as differential expression analysis (DEA), gene set enrichment analysis (GSEA) and cluster analysis.
- Sequential analysis methods, such as PCA, CA and NSCA, should be applied to all available genes or proteins, as well as the union of the top 50 differentially expressed genes or proteins, of the first data set from the HG-U133plus 2.0 platform and protein expression data of the NCI-60 data.
- The integrative analysis method MCIA should be applied to all available genes or proteins, as well as the union of the top 50 differentially expressed genes or proteins, of the first data set from the HG-U133plus 2.0 platform and protein expression data of the NCI-60 data.
- Finally, the analysis of sequential (PCA, CA; NSCA) and integrative (MCIA) methods should be compared.

2 Methods

2.1 Data sets

The data set used in this study is the US National Institute (NCI) data set comprising 60 human cancer cell lines (NCI-60). Therefore a brief overview of the data set will be given.

2.1.1 NCI-60 data set

This data represents gene and protein expression from a panel of 60 tumor cell lines (NCI-60) from the National Cancer Institute (NCI). These 60 human cancer cell lines consist of 9 different cancer tissue types, including breast, central nervous system (CNS), colon, leukemia, melanoma, non-small cell lung, ovarian, prostate and renal [86].

The gene expression data is a set of microarray gene expression profiles of the NCI-60 panel [86]. The already normalized data sets, as well as the raw Affymetrix data files (CEL files; Cell intensity file), are available through Cellminer [123,124] and can be downloaded from the Cellminer webpage [125]. These data come from four microarray platforms: Aligent GE 4x44K [44], Affymetrix HG-U95, Affymetrix HG-U133 and Affymetrix HG-U133 plus 2.0 technologies [42,43].

The Affymetrix human genome U95 set (HG-U95) consists of five Genechip arrays (HG-U95A – HG-U95E) [126–128]. This set includes about 63,000 probe sets which contains more than 60,000 covered transcripts [126–128]. The Affymetrix human genome U133 set (HG-U133) consists of two Genechip arrays (HG-U133A and HG-U133B) [126,129,130]. This set contains nearly 45,000 probe sets which represent approximately 39,000 transcript variants which again characterize about 33,000 known human genes [126,129,130]. The Affymetrix human genome U133 plus 2.0 (HG-U133 plus 2.0) is a single array which includes the complete HG-U133 set and in addition 6,500 new genes for the investigation of more than 47,000 transcripts [131,132]. These 6,500 additional genes from the array are represented by 9,921 new probe sets [131,132]. The last microarray platform is the Agilent-014850 Whole Human Genome Microarray 4x44K G4112F (feature number version) [133,134]. This microarray shows a composite view of the whole human genome and this Agilent one color microarray consists of four arrays spotted on each slide [133,134]. An overview of the processed raw files from the used microarray platforms can be found in Table 2. Another important aspect is that for each

platform the entire data set (NCI-60 data set) was measured. Over the years this data set was measured several times on different platforms (Table 2).

The CEL files contain the raw probe-level microarray data which are generated at the end of an array scan. Those files contain measured intensity and position values corresponding to each probe (probe ID) of a hybridized microarray. The raw CEL files are fed into R with the function *ReadAffy()* [135] which creates an AffyBatch object of the data [136]. Then the data sets are normalized with the function *gcrma()* [137]. This function converts the AffyBatch object into an expression set [136]. Finally, filtering was performed which, among other things, was responsible for removing the control probes, duplicated probes, as well as probes that do not belong to an official HUGO gene symbol. The raw data files of the Agilent platform consist of text files with data information and PDF files with quality information and are already processed and normalized using GeneSpring [138].

Initially the raw CEL files from the three different Affymetrix platforms were downloaded from Cellminer. The already normalized (GC robust multichip averaging (GCRMA) [139]) Affymetrix microarray data were downloaded from Cellminer as well. The Agilent data were only downloaded in the normalized version from Cellminer which are already log transformed and the control probes are removed [140].

The protein expression data of the NCI-60 panel was measured using mass spectrometry. The protein expression profiles from the individual cancer cell lines are generated by a conventional one-dimensional polyacrylamide gel electrophoresis (PAGE) and a subsequent in-gel tryptic digestion followed by liquid chromatography-tandem mass spectrometry (GeLC-MS/MS) and label-free quantification, as described by Gholami *et al.* [141,142]. The raw data files were processed with MaxQuant [143] where the data is normalized. The data sets can be downloaded as raw files, as well as in the normalized version, from the webpage of the Munich University of Technology [144]. In this study the already processed and normalized data set with label-free quantification was downloaded from the webpage of the Munich University of Technology and used for further analysis. An alternative web page which hold the raw data files is ProteomicsDB [145].

Table 2: Description of different microarray platforms

Platform	Technology	Manufacturer	Chips per platform
Human Genome U95	*in situ* oligonucleotide	Affymetrix	5
Human Genome U133	*in situ* oligonucleotide	Affymetrix	2
Human Genome U133 Plus 2.0	*in situ* oligonucleotide	Affymetrix	1
Whole Human Genome Microarray, 4 x 44K G4112F	*in situ* oligonucleotide	Agilent	1

Platform	Probe sets	Duplicated control probe sets	Overlap of probe sets between different chips per platform	Covered Transcripts
Human Genome U95	62,907	268	0	> 60,000
Human Genome U133	44,928	68	100	> 39,000
Human Genome U133 Plus 2.0	54,675	0	0	> 47,000
Whole Human Genome Microarray, 4 x 44K G4112F	45,220	0	0	~ 41,000

An overview of the technical replicates, total chips, total/used cell lines and scan dates of the different platforms can be found in Table 3. Therefore the scan date represents the hybridization date.

Table 3: Technical, as well as biological, facts and date of measurement of different microarray platforms

Platform	Technical replicates	Total chips/runs	Total cell lines	Used cell lines	Scan/measurement date
Human Genome U95	1	300	60	57	June – July 2001
Human Genome U133	1	118	59	57	April 2002
Human Genome U133 Plus 2.0	3	174	59, 59, 56	57, 57, 54	March – April 2005
Whole Human Genome Microarray, 4 x 44K	2 or 4	36	60	57	February 2007
Protein expression	1 – 2	746	59	57	September 2010 – June 2012

The raw data files of the Affymetrix and Agilent platforms were opened with the free source code editor Notepad++ 6.9.2 [146]. In the header of these data files the scan/measurement date can be checked. For the measurement date of the raw data files from the proteome data set we used the tool msconvert [147]. This tool allows the conversion from the raw proteome data files to mzML files. In these output files there is a row which contains the scan/measurement date of the proteome data set.

2.2 Re-annotation and Re-analysis

The Bioconductor [148] packages, such as hgu95a.db [149], hgu95b.db [150], hgu95c.db [151], hgu95d.db [152] and hgu95e.db [153], hgu133a.db [154], hgu133b.db [155] and hgu133plus2.db [156] are used for the annotation of the re-analyzed Affymetrix files. The Affymetrix data were normalized by GCRMA which considers the correction of the GC content of oligonucleotides [139]. The already normalized Affymetrix microarray data from the three different platforms

were downloaded from Cellminer. The original data files are merged by the probe IDs with the annotation files from the re-analyzed Affymetrix files. The annotation file for the Agilent 4x44K array was downloaded from the webpage of Gene Expression Omnibus [157]. The already normalized Agilent platform is downloaded from Cellminer and then merged by probe IDs with the annotation file from the known web page. In this annotation process the individual probes (probe IDs) are assigned to official HUGO gene symbols, the Entrez gene IDs and the gene ontology (GO) IDs. Here it should be mentioned if a probe ID (probe) map to a GO ID, it often maps to several GO IDs. In such a case only a single (the first) GO ID is selected.

A further step of re-analysis involves the comparison of intensity differences from the three Affymetrix platforms. Based on this step it will be decided which microarray platform is to be selected for the further processing and analysis steps since it differs the least from the original table. The differences of the expression data in all three histograms from the three different platforms (Figure 5) were generated by the subtraction of logarithmized intensity values from the original data table with the logarithmized intensity values from the re-annotated data table. The expression data of the re-annotated data table from the HG-U133 plus 2.0 platform were generated from three replicates. Therefore, the mean value of expression levels of the original data table from the HG-U133 plus 2.0 platform was calculated.

2.3 Differential expression analysis

An important aspect of microarrays is the measurement of mRNA expression levels corresponding to various genes. During the past decade microarrays have been widely used in transcriptomic studies, especially for the determination of differentially expressed (DE) genes under different conditions. Many different software packages have been developed for the analysis of the resulting data. Limma [158–160] is a software tool which is used for such analyses.

2.3.1 Computational steps of the differential expression analysis

Limma is a very popular and useful tool in R which enables the implementation of DEA of data from microarray and proteome experiments [158–160]. In our study the microarray expression data used for DEA are log-intensity values arising from the studies of one-channel Affymetrix HG-U133 plus 2.0 platform hybridized with one RNA sample to each array. Therefore we analyzed the microarray results from all three data sets of the HG-U133 plus 2.0 platform which considers the comparisons of many mRNA levels simultaneously between all pair-wise tissue

comparisons across the nine different tissues (Breast, CNS, Colon, Leukemia, Melanoma, Lung, Ovarian, Prostate and Renal). The three data sets of the proteome experiments are analyzed in the same manner as the microarray data.

The DEA of microarray and proteome data in limma is well described, implemented and explained in the following websites [160, 161–163]. Further details of the exact process of the DEA in limma, as well as some examples regarding DEA can be found in the same websites as mentioned above [160, 161–163]. When performing the DEA, we basically follow the workflow from these websites. The R-package limma under the version limma_3.28.6 was used for the calculations and analysis of DEA of the gene and protein expression data from the NCI-60 panel used [164].

A brief overview of DEA is described in the following paragraph and the information content was taken from the following websites [160, 161–163]. The basic principle of the DEA is to fit a linear model to the expression levels for all single genes. The first step in the DEA is to build a design matrix with the function *model.matrix()* [164]. This design matrix show which of the nine tissues have been applied to each array. Therefore we choose nine parameters in the linear model because we have nine different tissues in the microarray and proteome data. The nine columns of the design matrix represent the nine different tissues which correspond to the coefficients of the linear model. The rows of the design matrix denote the arrays in the experiment. Next, the expression and proteome data are fitted by the linear model with the R-function *lmFit()* [164] that models the entire systematic part of the microarray and proteome data. The fitted coefficients from the fitted model correspond to the mean-log expression values for each gene in all possible tissue combinations. Then the contrast matrix was created. The generated contrasts allow all possible pair-wise tissue comparisons across the nine different tissues. Therefore 36 tissue comparisons (all possible comparisons) between the nine different tissue combinations were generated in order to answer all questions of interest. The contrasts (columns of contrast matrix) correspond to the linear combinations of parameters from the linear model fit. The contrast matrix contains one row for all individual tissues and one column for all individual contrasts. Afterwards we determine the estimated contrasts with the R-function *contrast.fit()* [164] where we fit the contrast matrix to the above-mentioned fitted model. This function enables that the fitted coefficients can be compared in any way, so that all questions can be answered. The R-function *eBayes()* determines the consensus pooled variance [162]. Furthermore this function defines the empirical Bayes (moderated) pooled variance for all individual genes with the help of the previously computed variance [162]. This computation even helps for the adaption of the degrees of freedom for the contrast t-test [162]. The function

also determines the results for the t-test and the corresponding p-values [162]. More exact mathematical details regarding the linear model and all further steps which are involved in DEA can be found in the following websites [158,160,161–163,165].

The R-function *topTable()* [164] yields to a list of genes DE across all pair-wise comparisons between the nine different tissues of the three data sets from the HG-U133 plus 2.0 platform and the proteome data (Table 14). This function provides the following results: row numbers of probe IDs or HUGO gene symbols, IDs of HUGO gene symbols, absolute $\log_2$-fold changes between two experimental conditions (estimated contrasts) of different tissue comparisons, mean $\log_2$-expression values for each gene over the entire arrays, moderated t-statistics, p-value, adjusted p-value and the B-statistics which denotes the estimated log-odds probability ratios that the gene is DE [161,162]. In this case the most common method for significance analysis is the moderated t-statistics. This statistics is determined for all probes and contrast of the three data sets from the HG-U133 plus 2.0 platform and the proteome data. With the parameter *"coef"* in the *topTable()* function you can select the different contrast (tissue comparisons) in the data table. Furthermore, the parameter *"adjust.method"* in the *topTable()* function enables the selection of the adjustment method for the adjusted p-value. The adjustment method *"BH"* was selected which corresponds to the Benjamini and Hochberg's method [166]. It is one of the most common adjustment methods which control the false discovery rate (FDR) which was used to correct the p-values and genes/proteins with FDR ≤ 0.05 were recognized as significantly DE genes/proteins [166]. The number of over-expressed and under-expressed genes/proteins is determined based on the p-value (≤ 0.05) and the positive/negative $\log_2$-fold change (Table 14). The parameter *"lfc"* in the *topTable()* function enable the selection of the different cut-offs of the absolute $\log_2$-fold change values (1) wherein DE genes/proteins are selected with respect to this threshold (Table 14). The parameter *"genelist"* in the *topTable()* function allow the valid assignment to the already identified HUGO gene symbols of the probe sets.

2.4 Clustering and Heatmap

Cluster analysis [167–169] and heatmaps [167,168,170] are very common and useful tools which are applied on data from microarray experiments, as well as data arising from other high-throughput sequencing technologies. Therefore heatmaps enable the visual representation of these data. Additionally heatmaps can implement various clustering approaches which can be used simultaneously in rows (genes) and columns (arrays/samples) of the data matrices. These different clustering methods with various algorithms both allow the detection of clusters

regarding similarities in patterns of gene or protein expression data and the representation of these data in the same visual display. These clusters should show that observations (samples) in the same clusters are similar to each other compared to observations (samples) in other clusters.

In general cluster analysis can be divided into the two most important groups: hierarchical and partition-based clustering depending on how to arrange the data [171,172]. Hierarchical clustering groups the clusters into smaller ones in order to generate tree-shaped structures or dendrograms [172]. Compared to hierarchical clustering, partition-based clustering defines a predetermined number of sub-groups to which the data are assigned, without taking into account a hierarchical relation among the clusters [172]. The two sub-classes of hierarchical clustering are agglomerative and divisive clustering approaches [171–173]. Agglomerative hierarchical cluster analysis begins with individual observation (sample) clusters and then gradually closes the next clusters together until finally all observations (samples) are assigned to the supercluster [172]. Divisive hierarchical clustering is exactly the reverse method to agglomerative clustering and indeed divisive clustering starts with all observations (samples) in one cluster and then the cluster is subdivided into smaller groups until all observations (samples) join single observation (sample) clusters [171]. In this study we used just the above-mentioned cluster analysis method, called agglomerative hierarchical clustering.

2.4.1 Computational steps of clustering and heatmaps

The goal in our cluster analysis was to detect gene or protein expression patterns of three available data sets from the HG-U133 plus 2.0 platform and the proteome data. Therefore the hierarchical cluster analysis was used on Pearson correlation coefficients between the nine different tissues, as well as the 57 different cancer cell lines (samples). The aim therefore was to identify cluster between the nine different tissues and to find sub-groups in the 57 different samples where the samples in the same cluster were similar to each other.

First, the mean value of the gene and protein expression data of the different samples was calculated that belong to the same tissue origin with the R-function *avearray()* [164]. After that the Pearson correlation matrix was built for the nine different tissues and the 57 samples of the three data sets from the HG-U133 plus 2.0 platform and proteome data with the R-function *cor()* [174] where the argument “method” was set to “pearson”. The Pearson correlation coefficients were rounded to three decimal places.

Cluster analysis runs in a two-step process. Before the generation of the heatmap, the genes (rows) and/or arrays (columns) should be clustered in order to describe the similarity between samples. First, the distance between the different samples was calculated with the R-function *as.dist()* [174] where we use the Pearson correlation (centered correlation) [172] for the distance measurement. There are other distance measure methods, such as Manhattan distance, Euclidean distance, Mahalanobis distance, uncentered correlation, Spellman rank correlation and absolute or squared correlation, which can be read in [172]. Secondly, we should use one of the various tree construction approaches which differ due to the choice of the inter-cluster distance (cluster linkage methods) [172]. There are various methods for selection of cluster linkage methods, such as single linkage, complete linkage, average linkage and centroid linkage, can be found in [172]. Hierarchical clustering of rows is applied using the R-function *hclust()* [174] where we set the argument "method" to "average" which defines the average linkage clustering for the tree construction method. This linkage method, also known as unweighted pair-group method using arithmetic averages (UPGMA), defines the average distance among any two components [172]. The different heatmaps were generated with the R-function *heatmap.2()* [175]. Instead of using the default colors in *the heatmap.2()* function we use the R-function *colorRampPalette* [176] of the *RColorBrewer* package [176] in order to generate the own color palette from green over yellow to red. The argument "n = 50" in this function determines the individual colors in each color palette. The *heatmap.2()* function generates heatmaps where the row dendrograms (row clustering) have been added to the plot with the two arguments "dendrogram = row" and "Rowv". The "scale" argument retains the default value "rows" which means that the data matrix is scaled by rows. The argument "breaks" defines the color limits (from -1 to 0; from 0.1 to 0.8; from 0.81 to 1). The argument "color" introduces the color function in the heatmaps which represents low correlation coefficients (from -1 to 0) in different green tones, mean correlation coefficients (from 0.1 to 0.8) in different yellow tones and high correlation coefficients (from 0.81 to 1) in different red tones. The numeric values of the Pearson correlation coefficients in each cell can be represented by the argument "cellnote" where the argument "notecol" defines the color of the numeric values. The argument "trace = none" suppresses tracelines in the plot. The argument "RowSideColors" enables to annotate the heatmap with a color bar for row labels. The argument "key" represents a color key for the heatmap. The argument "density.info = none" suppresses the density plot in the color legend. The R-function *legend()* [177] adds a legend for the row labels to the heatmap. A detailed overview and application of the above-mentioned R-functions can be found in [173].

In the next step, the 57 cancer cell lines according to gene and protein expression data of the three data sets from the HG-U133 plus 2.0 platform and proteome data should be represented as dendrograms under the use of hierarchical clustering. For the hierarchical clustering of the 57 samples according to gene and protein expression data the data matrix had to be transposed first. The similarity between the clusters is described with the Euclidean distance measure [172] which is defined in the R-function *dist()* [174] by the argument "method = euclidean". The average linkage clustering is also used for the tree construction method in the R-function *hclust()* [174] in which the argument "method" was set to "average".

2.5 Enrichment analysis

The outcomes of the analysis of omics high-throughput data result frequently in a large list of interesting biomolecules, particularly large gene lists. The biological interpretation of large gene lists continues to be a challenge because it is difficult to determine their influence on abnormal biological processes, such as diseases and disorders, or their impacts on biological processes across a study, as well as many other questions. Therefore, the comprehensive understanding of the functional meaning of these gene lists is a difficult task.

The gene enrichment analysis is one of the most commonly used methods which link ontologies, pathways, and transcription factors to gene lists which are normally obtained from high-throughput experiments. For the performance of an enrichment analysis the following data are needed:

- A list of genes of interest which should be analyzed (subset of genes chosen by some relevant method)
- A reference list of gene groups (test the associated enrichment of the interested set of genes)
- A list of biological annotations, each linked to genes (GO terms)

A biological process is usually based on a group of genes rather than a single individual gene. The general principle of enrichment analysis relies on abnormal biological processes across studies where the co-functional genes should show a higher (enriched) potential. Therefore they should be recognized by high-throughput screening methods as relevant groups. Thus, the view of the analysis of large gene lists changes from a single individual gene-based analysis to a relevant group-based analysis [178].

The general principle of an enrichment analysis is based on a computational algorithm which investigates if a priori defined subset of genes reveals statistically significant or concordant differences among two biological states. Firstly, systematic map genes and proteins to their biological annotation, for example the corresponding GO terms [179]. After that, the comparison of the distribution of the terms in the group of genes, which has to be analyzed, with the background distribution of those terms, such as the total set of genes on a microarray chip, follows [180]. In such a case enrichment analysis enables to determine terms that are statistically over or under represented in the gene list which has to be analyzed [180,181]. From this it follows that these enriched terms characterize a few significant underlying biological processes [180].

Several enrichment methods for the analysis of large gene lists have been developed, such as Onto-Express [182], MAPPFinder [183], GoMiner [184], DAVID [185], EASE [186], GARBAN [187], GeneMerge [188] and FuncAssociate [189]. After these incipient tools more public available enrichment analysis methods follow. Khatri *et al.* [190] and Curtis *et al.* [191] summarized and published about 14 other enrichment tools in the year 2005. Furthermore, Huang *et al.* published a survey and summary about several bioinformatics enrichment tools [178]. In this study, about 68 different bioinformatics enrichment tools are presented and described in detail. Huang *et al.* divided the variety of enrichment analysis methods in three different classes, with respect to the differences in the enrichment algorithms: singular enrichment analysis (SEA), gene set enrichment analysis (GSEA) [192,193] and modular enrichment analysis (MEA) [178].

2.5.1 Computational steps of the gene set enrichment analysis

Before the performance of the GSEA of GO categories and Reactome pathways of gene expression data, microarray probe IDs should be converted to Entrez IDs. NCBI employs these Entrez IDs in order to connect various databases. This allows that only the unique Entrez IDs from the genes of interest were obtained. Different annotation packages available for various microarray platforms in R allow the conversion from probe IDs of microarray platforms to Entrez IDs. The matrix from the genes of interest arose from the Affymetrix HG-U133 plus 2.0 platform and therefore we used the suitable annotation package hgu133plus2.db.

Before the performance of the GSEA of GO categories and Reactome pathways of protein expression data, the HUGO gene symbols of the available IPI protein identifiers were mapped to the HUGO gene symbols of the HG-U133 plus 2.0 platform. All gene symbols which were not

found in the HG-U133 plus 2.0 platform were removed. Then these HUGO gene symbols, which were obtained from the HG-U133 plus 2.0 platform, were converted to their corresponding Affymetrix probe IDs. Afterwards these Affymetrix probe IDs should be converted to Entrez IDs. The used annotation package for the GSEA of the protein expression data was also the annotation package hgu133plus2.db.

A detailed overview and application of the GSEA method and their associated R-functions can be found in [194]. In principle, the workflow for the GSEA in [194] was kept. The R-function *new()* [195] from the R-package category [195] was used to define the parameter for the GSEA and therefore the first argument of this function was set to "GOHyperParams". The R-function *hyperGTest()* [196] from the R-package GOstats [196] was used for the GSEA for GO categories. This GSEA was applied to statistically significant DE genes/proteins which were obtained by the analysis with limma as mentioned in section 2.5.1.

The GSEA for the GO categories is a three-step process because the GO hierarchy is composed of three different ontology categories: biological process (BP), molecular function (MF) and cellular component (CC) [197,198]. The parameters for the hypergeometric test [199,200] are defined in the R-function *new()* where the arguments for this function are described as follows: *"geneIds"* which denotes the Entrez IDs of the genes/proteins from interest (DE genes/proteins), *"universeGeneIds"* which characterizes the Entrez IDs from the entire HG-U133 plus 2.0 platform, *"annotation"* which defines the suitable annotation package for the data, *"ontology"* which describes the ontology category (BP, MF and CC), *"pvaluecutoff"* which determines the p-value cutoff and *"testDirection"* which characterizes over –or under-enrichment direction. With this parameters the computation of the hypergeometric test can be started with the R-function *hyperGTest()*. We chose the p-value cutoff 0.05 which proves how many GO terms are statistically significant under the selection of this cutoff. For the parameter *"testDirection"* we selected the argument *"over"* which uses testing for over-enrichment analysis.

The R-function *enrichPathway()* [201] from the R-package ReactomePA [201] was used for the GSEA for Reactome pathways. This GSEA was applied to statistically significant DE genes/proteins which were obtained by the analysis with limma as mentioned in section 2.5.1. The parameters for the GSEA for the Reactome Pathways are defined in the R-function *enrichPathway()* where the arguments for this function are described as follows: *"gene"* which denotes the Entrez IDs of the genes/proteins from interest (DE genes/proteins), *"universe"* which characterizes the Entrez IDs from the entire HG-U133 plus 2.0 platform, *"pvaluecutoff"* which determines the p-value cutoff and *"organism"* which characterizes the existing organism.

With this parameters the computation of the GSEA for the Reactome pathways can be started with the R-function *enrichPathway()*. We chose the p-value cutoff 0.05 which proves how many Reactome pathways are statistically significant under the selection of this cutoff.

2.6 Sequential analysis methods

Sequential analysis methods, including DRTs for one data set [88], are commonly used tools for the analysis of multi-omics data. These DRTs are the most mentioned and important methods which are counted among the sequential analysis methods. These DTRs arose during the early 20^{th} century [202]. The goal of DRTs is to decompose the data table (matrix) into a smaller set of novel variables (components) with the aim to detect as much differences as possible in the different samples [88]. PCA [88,100–103], CA [104–107] and NSCA [106,108,109] will be used as sequential analysis methods in this study. The high-dimensionality makes it difficult to extract the valuable information of the entire datasets. Therefore, a DRT should be used to reduce the dimensionality of the data sets to generate a subset of suitable dimensions in order to pool and extract the most relevant variant features of the data sets and to obtain as much information as possible.

The ade4 [203,204], FactoMineR [205] and stats R packages enable the use and implementation of PCA, as well as many other multivariate statistical analysis methods. CA is available in R packages, for instance ca [206], FactoMineR [205] and ade4 [203,204]. NSCA is implemented in the R package ade4 [203,204].

In the following matrices are denoted in bold upper-case letters, vectors in bold lower-case letters and scalars in lower-case letters. The superscript T defines the transposition.

A typical omics data table $\boldsymbol{X}$ consists of I measurements (observations) and J features (variables), whereby the measurements (observations) and features (variables) are represented as rows and columns of the matrix $\boldsymbol{X}$. The features (variables) correspond to typical cellular components in a biological system (cell, tissue or organism), such as mRNA, protein and metabolite levels.

The mathematical concept, the corresponding mathematical equations and its description are summarized from Meng *et al.* [88].

The available omics data table $\boldsymbol{X}$ is an I x J matrix, with the elements x_{ij}, where $i = 1, \dots, I$ and $j = 1, \dots, J$. The matrix $\boldsymbol{X}$ consists of I measurements (observations) and J features (variables). It can be expressed by equation 1:

$$\boldsymbol{X} = (x_1, x_2, \ldots, x_j) \quad (1)$$

The $\boldsymbol{x}_j$*'s* denote vectors with the length of *I*. These vectors contain mRNA or protein expression data, as well as other biological features (variables) for *I* measurements (observations). In characteristic omics study, the index *J* varies between a few hundreds and million values. Therefore these measurements (observations) are shown in a high-dimensional space $\mathbb{R}^J$.

2.6.1 Principal component analysis

PCA is a useful statistical exploratory technique for the analysis of multivariate data [207,208]. It is one of the earliest and most widespread dimension reductions techniques. PCA has been applied to various research fields including biomedical studies, such as the investigation of microarray data regarding identification of outlier genes [209]. Moreover PCA was also successfully applied for studying other types of expression data [210]. PCA can determine patterns in data sets, as well as graphically represent similarities and differences of data. When the biological question is associated with the highest variance, then PCA will be an efficient tool for the investigation of this biological question. Furthermore PCA is frequently used as a pre-processing step in order to reduce the dimensionality in complex data sets [211,212].

The main aims of PCA are [213]:

- to detect the most important information from the data
- to reduce the dimensionality in order to obtain only the important information
- to facilitate the interpretation of the data sets
- to analyze the structure of the observations (measurements) and variables (features)

2.6.2 Mathematical concept of principal component analysis

The mathematical concept, the corresponding mathematical equations and its description are summarized from Meng *et al.* [88].

As the variance depends on the scale of variables, all column vectors $\boldsymbol{x}_j$ are standardized in order to obtain a mean value of zero and a standard deviation of 1 [214]. After this, all original variables have unit norm and can be compared [214].

PCA identifies novel variables $\boldsymbol{f}^{i} = \boldsymbol{X}\boldsymbol{\omega}^{i}$, in which i denotes the i^{th} component. The index $\boldsymbol{\omega}^{i}$ stands for the variable loading (components) for the i^{th} principal component (PC). The exponent expresses the component or the dimension. The maximization of the variance of $\boldsymbol{f}^{i}$ can be expressed by equation 2:

$$arg\,max_{q^i}\ var\left(\boldsymbol{X}\boldsymbol{\omega}^{i}\right) \tag{2}$$

This maximization of the variance is restricted by the condition $\left\|\boldsymbol{\omega}^{i}\right\| = 1$. Any novel pair of components $(\boldsymbol{f}^{i},\boldsymbol{f}^{j})$ is orthogonal to one another or uncorrelated which means that the result of the calculated correlation between each pair of components will be zero, for instance $\boldsymbol{f}^{iT}\boldsymbol{f}^{j} = 0$ for any condition $j \neq i$.

The PCA can be calculated using singular value decomposition (SVD) [88,215,216]. PCA can use also many other mathematical algorithms, but SVD is the most common one. The matrix $\boldsymbol{X}$ (I x J matrix), with the rank $r \leq \min(I,J)$ of the matrix $\boldsymbol{X}$ which is defined as the largest number of linearly independent column or row vectors (both definitions are equivalent), can be computed by SVD, as indicated by equation 3:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{Q}^{T}\ subject\ to\ the\ constraint\ \boldsymbol{U}^{T}\boldsymbol{U} = \boldsymbol{Q}^{T}\boldsymbol{Q} = \boldsymbol{1} \tag{3}$$

SVD decomposes $\boldsymbol{X}$ into three matrices depending on the orthonormalizing constraints $\boldsymbol{U}^{T}\boldsymbol{U} = \boldsymbol{Q}^{T}\boldsymbol{Q} = \boldsymbol{1}$. $\boldsymbol{U}$ denotes the I x r matrix, where the columns of $\boldsymbol{U}$ are called left singular vectors and $\boldsymbol{Q}$ denotes the J x r matrix, where the columns of $\boldsymbol{Q}$ are the right singular vectors. $\boldsymbol{S}$ denotes the r x r diagonal matrix of singular values where the singular values are sorted in descending order. In the case of the PCA of $\boldsymbol{X}$, the PCs consist of a (I x r) matrix, $\boldsymbol{F}$. The matrix $\boldsymbol{F}$ is expressed by the following equation (equation 4):

$$\boldsymbol{F} = \boldsymbol{U}\boldsymbol{S} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{Q}^{T}\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{Q} \tag{4}$$

The PCs are represented by the columns of the matrix $\boldsymbol{F}$. The matrix $\boldsymbol{Q}$, also termed as loading matrix, is defined by the coefficients of the linear combinations of the variables for any PC. Hence, the variance of $\boldsymbol{X}$ is displayed by a lower dimensional space r. Equation 4 reveals that $\boldsymbol{Q}$ is a matrix that projects measurements (observations) into $\boldsymbol{X}$ on the PCs.

The PCs capture as much of the variance in the original variables as possible [213]. The first PC extracted in PCA detects the largest variance [213]. The second largest variance corresponds to the second PC that is orthogonal to the first PC [213]. More PCs are determined in the same manner [213]. However, it is common to use the first few PCs which detect reasonable amounts

of variance in the original datasets because the last few PCs are often associated with the residual noise in the data [102].

The variance of the i^{th} PC (d^{i^2}) can be calculated by the following equation (equation 5) because the column sum of the squared values in $\boldsymbol{U}$ equals to one (equation 3) [88]:

$$d^{i^2} = \frac{s^{i^2}}{n-1} \tag{5}$$

In equation 5 the i^{th} diagonal element in $\boldsymbol{S}$ denotes s^i. The variance represents the quantity of information which is detected by any PC [88].

2.6.3 Computational steps of the principal component analysis

The principal component analysis should be applied to the 57 cancer cell lines according to gene and protein expression data of the three data sets from the HG-U133 plus 2.0 platform and proteome data. First, the data frame had to be converted into a data matrix and then this data matrix should be transposed. Afterwards the principal component analysis was performed with the R-function *prcomp()* [174] using the R-package *stats* [174]. In the plotted result figure of the PCA, a legend can be added for the labels in the plot using the R-function *legend()* [177]. A detailed overview and application of the above-mentioned R-functions can be found in [173].

The PCA of the union of the top 50 DE genes or proteins of the three different data sets from the HG-U133 plus 2.0 platform and proteome data was performed in the same manner as the PCA from all available genes or proteins. But in this case, only the top 50 DE genes or proteins were primarly selected from all the existing 36 tissue comparisons, and then the union of these genes or proteins was formed.

Finally, the most important genes at the positive and negative end of the first and second principal component were determined with the R-function *topgenes()*.

2.6.4 Correspondence analysis

CA is a statistic exploratory method applied to various research areas which visualizes the relationships between rows and columns in data tables [217]. The number of dimensions is reduced in order to enable the graphical representation of the data points [218]. In particular this is achieved by decomposing the total inertia (i.e. variability) of the data table and determines the smallest amount of dimensions that can detect the data variability [218]. The resulting graphical

representation of the datasets facilitates the interpretation of the data. This visualization of the data table is represented as a set of two-dimensional scatterplots where the samples and/or features are projected as points onto a lower-dimensional space [218]. The CAinterprTools R package is a tool which enables a graphic support in order to facilitate the interpretation of the results [218].

2.6.5 Mathematical concept of correspondence analysis

Similar to PCA, CA uses SVD for dimension reduction, but relies on a different transformation of the dataset before decomposition [107,216].

The goal of CA is to project rows, as well as columns of a data table in the same dimensional space [105]. Particularly this is done for the first few coordinates (two or three) which contain the highest amount of the relevant information [105]. The mathematical concept, the corresponding mathematical equations and its description are taken from Fellenberg *et al.* and Greenacre *et al.* [105,107]. Further details and information regarding this method are described in [104,107].

The input data table is a I x J matrix $\boldsymbol{X}$, with the elements x_{ij}, where $i = 1, \ldots, I$ and $j = 1, \ldots, J$. It consists of I measurements (observations), which correspond to the number of probes or genes from the four different microarray platforms, and J features (variables), which correspond to the 57 cancer cell lines from the four different microarray platforms. I denotes the rows and J the columns of the data table $\boldsymbol{X}$. Further x_{i+} and x_{+j} represent the row and column sums (totals) of $\boldsymbol{X}$ and x_{++} corresponding to the grand total of $\boldsymbol{X}$ [105]. The weight of the j^{th} column (column masses) is defined by the following equation (equation 6) and these column masses are assigned to the column profiles [105,107]:

$$c_j = \frac{x_{+j}}{x_{++}} \tag{6}$$

The weight of the i^{th} row (row masses) is defined as shown (equation 7) and these row masses are assigned to the row profiles [105,107]:

$$r_i = \frac{x_{i+}}{x_{++}} \tag{7}$$

The correspondence matrix, or matrix of relative frequencies, $\boldsymbol{P} = [p_{ij}]$, which is the basis for the computation, is obtained by the following equation (equation 8) [105,106]:

$$p_{ij} = \frac{x_{ij}}{x_{++}} \tag{8}$$

Similar to PCA SVD is used in order to calculate coordinates of the row and column profiles in relation to principal axes [107].

The matrix $\boldsymbol{S} = [s_{ij}]$ of standardized residuals, which is derived from equation 8 can be calculated as described by Fellenberg *et al.* (equation 9) [105]:

$$s_{ij} = \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}} \tag{9}$$

$\boldsymbol{S}$ is subjected to SVD. This means that $\boldsymbol{S}$ is decomposed into three matrices $\boldsymbol{S} = \boldsymbol{U\Lambda Q}^T$ depending on the orthonormalizing constraints $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{Q}^T\boldsymbol{Q} = \mathbf{1}$ [88]. A detailed description of SVD can be found in section 2.6.2. $\boldsymbol{U}$ and $\boldsymbol{Q}$ denote the left and right singular vectors [107]. $\boldsymbol{\Lambda}$ stands for a diagonal matrix and its diagonal elements are denoted as the singular values of $\boldsymbol{S}$ which are sorted in descending order and are referred to as α_k: $\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_k$ [105].

Equation 10 represents the standard coordinates a_{ik} of rows [105,107]:

$$a_{ik} = \frac{u_{ik}}{\sqrt{r_i}} \tag{10}$$

The standard coordinates b_{jk} of columns are calculated as follows (equation 11) [105,107]:

$$b_{jk} = \frac{q_{jk}}{\sqrt{c_j}} \tag{11}$$

These standard coordinates (a_{ik}, b_{jk}) allow the representation of associations between row and column categories. They can be visualized as coordinates of the i^{th} row and j^{th} column onto the k^{th} dimension of the two-dimensional CA plot [106].

The coordinates of the row (measurement) i in the novel subspace can be calculated as follows (principal coordinates of rows) (equation 12) [105]:

$$f_{ik} = \frac{\alpha_k u_{ik}}{\sqrt{r_i}} \; with \; k = 1, .. \tag{12}$$

The row index i of the principal coordinates of rows f_{ik} denotes the rows of the original data table $\boldsymbol{X}$, while the column index k of the principal coordinates of rows f_{ik} denotes the principal axes or dimensions of the original data table $\boldsymbol{X}$.

For columns (features) in the same plane, the coordinates of the column (feature) j can be defined in the following manner (principal coordinates of columns) (equation 13) [105]:

$$g_{jk} = \frac{\alpha_k q_{jk}}{\sqrt{c_j}} \; with \; k = 1, .. \tag{13}$$

The row index j of the principal coordinates of columns g_{jk} denotes the columns of the original data table $\boldsymbol{X}$, while the column index k of the principal coordinates of columns g_{jk} denotes the principal axes or dimensions of the original data table $\boldsymbol{X}$.

These coordinates (f_{ik}, g_{jk}) are denoted as principal coordinates [105]. These two principal coordinates have the ability to reflect the strength of association that exists between two variables [106].

Similar to PCA, the aim of CA is to reduce the dimensionality of the data table and that is the reason why the data matrix is projected onto lower dimensions, mainly two or three [105]. Therefore, only a few (two or three) coordinates of the novel subspace are displayed.

The inertia Φ can be calculated by the sum of squares of the singular values which correspond to the sum of principal inertias (equation 14) [107]:

$$\Phi = \sum_{k=1}^{K} \alpha_k^2 = \sum_{k=1}^{K} \lambda_k \tag{14}$$

The loss of information is related to the dimension reduction and is quantified by the so-called total inertia Φ^2 which is the sum of squares of the principal inertias along each of the K dimensions (equation 15) [106]:

$$\Phi^2 = \sum_{k=1}^{K} \lambda_k^2 \; where \; K = min\{I - 1, J - 1\} \tag{15}$$

The chi-square (χ^2) statistic calculates the difference between all pairs of observed and expected frequencies in a contingency table. The differences are squared and finally they are divided by the expected frequencies. More details about χ^2 statistic and their calculations and can be taken from [107]. Equation 16 shows that the total inertia of a contingency table is the χ^2 statistic divided by the grand total of the data table $\boldsymbol{X}$ [107]:

$$\Phi^2 = \frac{\chi^2}{x_{++}} \tag{16}$$

The total inertia which characterizes the total variance of the data table is similar to the chi-square statistic but the calculation is based on relative observed and expected frequencies. The total inertia equals the sum of squares of the matrix $\boldsymbol{S}$ (see equation 9) (equation 17) [107,206]:

$$\Phi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(p_{ij}-r_i c_j\right)^2}{r_i c_j} \tag{17}$$

2.6.6 Computational steps of the correspondence analysis

There are three different packages in R, which contain functions for CA. We used the package *ade4* [219] with the R-function *dudi.coa()* [219] because the application of the other two packages, *FactoMineR*[1] [220] and *ca*[2] [221], with the functions *CA()*[11] [220] and *ca()*[12] [221] for the correspondence analysis caused an error message (error 1[11] and error 2[12]) in the first data set of the protein expression data in the R-software.

The correspondence analysis should be applied to the 57 cancer cell lines according to gene and protein expression data of the first data sets from the HG-U133 plus 2.0 platform and proteome data. First, the data frame had to be converted into a data matrix. Afterwards the correspondence analysis was performed with the R-function *dudi.coa()*.

The CA of the union of the top 50 DE genes or proteins of the first data sets from the HG-U133 plus 2.0 platform and proteome data was performed in the same manner as the CA from all available genes or proteins. But in this case, only the top 50 DE genes or proteins were primarly selected from all the existing 36 tissue comparisons, and then the union of these genes or proteins was formed.
Finally, the most important genes at the positive and negative end of the first and second dimension were determined with the R-function *topgenes()*.

[1] Error 1 in *FactoMineR* package with the R-function *CA()* of the first data set of proteome data:"Error in eigen(crossprod(X, X), symmetric = TRUE) : infinite or missing values in 'x' "

[2] Error 2 in *ca* package with the R-function *ca()* of the first data set of proteome data: "Error in svd(S) : infinite or missing values in 'x' "

2.6.7 Non-symmetric correspondence analysis

In the previous chapter, we focused on the analysis of the interdependence or symmetric dependence of (categorical) variables which describes the symmetric association structure among two variables of a contingency table [106]. This means that both variables are handled as predictor variables and none of them is regarded as a response variable [106]. This interdependence analysis structure mainly concerns the CA. In contrast, the analysis of the dependence, which describes the asymmetric association structure among two variables of a contingency table (predictor-response association), corresponds to one or more criterion variables (response variables), of which one can assume that they are dependent on other variables, also known as predictor variables [106,222]. PCA is a useful tool for the analysis of continuous variables, whereas NSCA is an exploratory technology which was developed for the analysis of dependence for a few variables (usually two or three) [109]. Additionally, it can also be extended to investigate several variables [109].

2.6.8 Mathematical concept of non-symmetric correspondence analysis

The mathematical concept, the corresponding mathematical equations and its description are summarized from Lombardo *et al.* [106].

The input data table is a I x J matrix $\boldsymbol{X}$, with the elements x_{ij}, where $i = 1, \dots, I$ and $j = 1, \dots, J$. It consists of I measurements (observations), which correspond to the number of probes or genes from the four different microarray platforms, and J features (variables), which correspond to the 57 cancer cell lines from the four different microarray platforms. I denotes the rows and J the columns of the data table $\boldsymbol{X}$.

Given a I x J contingency table $\boldsymbol{X}$, where $\mathbf{X}$ classifies two categorical variables. These two variables comprise I row and J column categories. Consider that there are two categorical variables where the column variable is referred to as categorical predictor variable, whereas the row variable characterizes the categorical response, or dependent, variable. The i^{th} and j^{th} relative frequency is indicated by p_{ij}. The i^{th} row marginal relative frequency is characterized by the following equation (equation 18):

$$p_{i+} = \sum_{j=1}^{J} p_{ij} \tag{18}$$

The j^{th} column marginal relative frequency is calculated by equation 19:

$$p_{+j} = \sum_{i=1}^{I} p_{ij} \tag{19}$$

The NSCA is based on the principle of the Goodman-Kruskal tau statistics. The proportional reduction of error in prediction which equals the increase in predictability regarding the response variable, pertaining to a predictor variable, can be calculated using the tau index (equation 20):

$$\tau = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} p_{+j}(\frac{p_{ij}}{p_{+j}} - p_i)^2}{1 - \sum_{i=1}^{I} p_i^2} = \frac{\tau_{num}}{1 - \sum_{i=1}^{I} p_i^2} \tag{20}$$

τ_{num} which is also referred to as numerator of τ defines the predictability of the response variable (rows) in relation to the predictor variable (columns). In contrast, the denominator defines the error in prediction. This error is independent of the predictor categories.

The numerator of τ, τ_{num}, is expressed in equation 21:

$$\tau_{num} = \sum_{i=1}^{J}\sum_{j=1}^{J} p_{+j}\,(\frac{p_{ij}}{p_{+j}} - p_i)^2 \tag{21}$$

The implementation of NSCA needs the decomposition of the Goodman-Kruskal tau index, τ, into the sum of squares of the singular values of the weighted centered column profile table. The centred profile $\boldsymbol{\Pi} = [n_{ij}]$ which characterize, how a column profile differs from a row marginal relative frequency, is represented using the matrix form in equation 22:

$$\boldsymbol{\Pi} = \frac{p_{ij}}{p_{+j}} - p_{i+} \tag{22}$$

The Goodman-Kruskal tau index is decomposed by generalized singular value decomposition (GSVD) to n_{ij} (equation 23) and expressed in matrix form $\boldsymbol{\Pi}$ (equation 24) in order to perform the NSCA [106]:

$$\pi_{ij} = \sum_{m=1}^{M} \lambda_m a_{im} b_{jm}, \tag{23}$$
$$with\ i = 1, \dots, I, j = 1, \dots, J\ and\ M = 1, \dots, min(I, J) - 1$$

$$\boldsymbol{\Pi} = \boldsymbol{A}\Lambda_{\lambda}\boldsymbol{B^T} \tag{24}$$

$\boldsymbol{\Lambda}_{\lambda}$ denotes a M x M diagonal matrix of singular values $\boldsymbol{\Pi}$ with elements n_{ij} [106]. $\boldsymbol{A}$ denotes a I x M column matrix comprising of I left singular vectors where the i^{th} and m^{th} component is a_{im} with $i = 1, \ldots, I$ and $m = 1, \ldots, M$ [106]. $\boldsymbol{B}$ denotes a J x M column matrix comprising of J right singular vectors where the j^{th} and m^{th} component is b_{jm} with $j = 1, \ldots, J$ and $m = 1, \ldots, M$ [106]. The conditions of the constraints can be found in [106].

τ_{num}, can also be denoted as the sum of squares of the singular values (equation 25):

$$\tau_{num} = \sum_{m=1}^{M} \lambda_m^2 \tag{25}$$

Similar to CA, which considers two symmetrically associated variables, there are also opportunities in NSCA to graphically represent the asymmetric association between variables [106]. Beh *et al.* mentioned three types: standard coordinates, principal coordinates and biplot coordinates [106].

2.6.9 Computational steps of the non-symmetric correspondence analysis

The NSCA should be applied to the 57 cancer cell lines according to gene and protein expression data of the first data sets from the HG-U133 plus 2.0 platform and proteome data. First, the data frame had to be converted into a data matrix and then the non-symmetrical correspondence analysis was performed with the R-function *dudi.nsc()* [219] using the R-package *ade4* [219].

The NSCA of the union of the top 50 DE genes or proteins of the first data sets from the HG-U133 plus 2.0 platform and proteome data was performed in the same manner as the NSCA from all available genes or proteins. But in this case, only the top 50 DE genes or proteins were primarly selected from all the existing 36 tissue comparisons, and then the union of these genes or proteins was formed.

Finally, the most important genes at the positive and negative end of the first and second dimension were determined with the R-function *topgenes()*.

2.7 Integrative analysis methods

Integrative analysis methods [223] are important tools for the analysis of multi-omics data. Compared to sequential analysis methods it has to be noted that the integrative analysis methods can be applied for the analysis of multiple (more than two) data sets. This can be seen as the

main advantage compared to sequential analysis methods, since they enable the simultaneous analysis of high-dimensional data sets much easier.

The fast progress in the development of high-throughput methods allows the measurement of large amounts of different biological molecules, such as DNA, mRNAs, proteins and metabolites, for omics approaches. This is the reason for an increased amount of information on the multi-level omics scale. Therefore multivariate methods are required which enable the integration and analysis of multi-level omics data in order to provide deeper insight into a biological system.

These multivariate technologies are very suitable for the analysis of multi-omics data. The omicade4 package [84–86] which has evolved from the two packages, ade4 [203,204] and made4 [224], facilitates the integration and analysis of multiple omics datasets. One common approach for the analysis of high-dimensional omics data is MCIA, which is an extended version of CIA.

2.7.1 Co-inertia analysis

CIA detects common trends or relationships in paired omics data sets [89] and deals with the determination of ordinates, also termed dimension reduction diagrams [112]. This is achieved by the determination of consecutive axes from two data sets in which the covariance between eigenvectors are maximized [89]. CIA is a two-step process: in the first step CIA can use DRTs, such as PCA, CA or NSCA, on the original data sets depending on different data types (binary, categorical, discrete counts or continuous data) [88,110–112]. The second step is the generalization of CIA which is based on a certain constraint criterion that constrains the projections of the orthogonal axes in the way that they show maximal covariance [88,110–112].

Thus, the first step of CIA corresponds to the separate analysis of two data tables to be analyzed [110]. It intends to detect axes which are maximizing the inertia in each hyperspace [110]. If all variables of a given data table are centered, then the total inertia of the covariance matrix results from the sum of the variances [110]. The second step is the generalization of CIA which underlies a certain constraint criterion. Therefore the goal of CIA is to determine a set of co-inertia axes (orthogonal vectors) in the two data sets on which the variables are projected [110]. One important principle of CIA is that it maximizes the squared covariance (co-inertia) among the projections of the variables on the co-inertia axes [110]. After that the squared covariance for this data can be decomposed. This CIA decomposition shows that the variance

and the correlation among two data tables are maximized [88]. Therefore, it is less sensitive to outliers [88]. Further details regarding CIA and its mathematical concept can be found in [88,110–112].

2.7.2 Multiple co-inertia analysis

MCIA, which is an extension of CIA, has the ability to analyze multiple data tables using a covariance optimization criterion [86]. This integrative analysis method seems to be very useful for pooling and isolating the most variant features through datasets, like genes and proteins, which are projected onto the same dimensional space [86]. MCIA is performed in two independent steps [86,88]: during the first step of MCIA dimension reduction, such as PCA, CA and NSCA, is applied to each data set in order to simplify the initial datasets. Therefore, the data are projected onto comparatively lower dimensions. More specifically, the first two dimensions are used in the further course. Step one is followed by a CIA analysis. This step focuses on the generalization of CIA which is based on a certain criterion that constrains the projections of the orthogonal axes in a way that they are maximally covariant [88,110–112].

The omicade4 R package [84–86] enables the use of MCIA and provides several functions for the visualization of MCIA outputs. The resulting graphical representation facilitates the interpretation of data.

2.7.3 Mathematical concept of multiple co-inertia analysis

The mathematical concept, the corresponding mathematical equations and its description are summarized from Meng *et al.* [86].

A typical omics data table consists of a n x m matrix $\boldsymbol{X} = [x_{ij}]$ including I measurements (observations) and J features (variables), with $1 \leq i \leq I$ and $1 \leq j \leq J$. The measurements (observations) and features (variables) are represented as rows and columns of the matrix $\boldsymbol{M}$. Further m_{i+} and m_{+j} denote the row and column sums of $\boldsymbol{M}$ and m_{++} corresponding to the grand total of $\boldsymbol{M}$. The relative contribution or weight of the row i to the total variation in the dataset (r_i) can be calculated with the following equation (equation 26):

$$r_i = \frac{m_{i+}}{m_{++}} \tag{26}$$

The relative contribution or weight of the column j to the total variation in the dataset (c_j) is defined by the following equation (equation 27):

$$c_j = \frac{m_{+j}}{m_{++}} \tag{27}$$

The weight of each individual element of $\boldsymbol{M}$ to the total variation in the dataset (p_{ij}) can be calculated by equation 28:

$$p_{ij} = \frac{m_{ij}}{m_{++}} \tag{28}$$

Then a novel matrix $\boldsymbol{M} = [m_{ij}]$ can be deduced by the above-mentioned factors (equation 29):

$$m_{ij} = \frac{p_{ij}}{r_i} - c_j \tag{29}$$

In equation 29 m_{ij} denotes the centered row profile, in other words the relative abundance of selected variables to the contribution of measurements.

The mathematical concept for the CIA analysis, the corresponding mathematical equations and its description are taken from Hua *et al.* [122]. Further details and information regarding this method and a precise mathematical description are reviewed in [86].

These techniques provide a solution for the simultaneous analysis of a number of statistical triplets ($\boldsymbol{M}_k$, $\boldsymbol{Q}_k$, $\boldsymbol{D}$) [86]. $\boldsymbol{M}_k$ denotes the range of transformed matrices. $\boldsymbol{Q}_k$ denotes the hyperspace of feature metrics [122]. $\boldsymbol{D}$ denotes an identity matrix [122]. During MCIA the sum of the squared covariance among the scores of any table with synthetic axes v is maximized (equation 30) [122]:

$$f(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k, \ldots, \boldsymbol{u}_K, \boldsymbol{v}) = \sum_{k=1}^{K} w_k \times cov^2(\boldsymbol{M_k Q_k u_k}, \boldsymbol{v}) \tag{30}$$

The term $cov^2(\boldsymbol{M_k Q_k u_k}, \boldsymbol{v})$ denotes the square of covariance and w_k is the weight of any table. The $\boldsymbol{u_k}$ represents auxiliary axles and $\boldsymbol{v}$ denotes the reference data structure or the synthetic center. $\boldsymbol{M}$ can be interlinked as described in equation 31 [122]:

$$\boldsymbol{M} = \left[w_1^{\frac{1}{2}} \boldsymbol{X_1} \,|\ldots|\, w_K^{\frac{1}{2}} \boldsymbol{X_K} \right] \tag{31}$$

Similarly, $\boldsymbol{Q}$ can be interlinked as described in equation 32 (equation 21) [122]:

$$\boldsymbol{Q} = [\boldsymbol{Q_1}|\ldots|\boldsymbol{Q_k}] \tag{32}$$

One can calculate the first order solution $\boldsymbol{u_i^1}$, with $1 \leq i \leq K$, and $\boldsymbol{v^1}$ through the first PC of the eigen-system by the following equation (equation 33) [122]:

$$w\boldsymbol{MQM^T Dv} = \lambda \boldsymbol{v} \tag{33}$$

The normalized auxiliary axis $\boldsymbol{u_k^1}$ is represented by:

$$\boldsymbol{u_k^1} = \frac{\boldsymbol{M_K^T D v^1}}{\left\| \boldsymbol{M_K^T D v^1} \right\|_{\boldsymbol{Q_k}}} (k = 1, 2, \dots, K) \tag{34}$$

The other solutions are obtained with the support of residual matrices from the calculation of the solutions of the first order, so that the remaining order axes being orthogonal with the previous set (equation 35, equation 36) [86]:

$$\boldsymbol{v^{jT} D v^s} = 0 \ (1 \ \leq j \ < s) \tag{35}$$

$$\boldsymbol{u_k^{jT} Q_k u_k^s} = 0 \ (1 \ \leq j \ < s) \tag{36}$$

The solution of the second order is obtained by the following equation (equation 37) [122]:

$$\boldsymbol{M_{1(order2)}} = \boldsymbol{M_1 - M_1 P_k^1} \tag{37}$$

The projecting matrix $\boldsymbol{P_k^1}$ is represented by equation 38 [122]:

$$\boldsymbol{P_k^1} = \boldsymbol{u_k^1 (u_k^1 Q_k u_k^{1T})^{-1} u_k^1 Q_k} \tag{38}$$

The two step process is performed until a meaningful number of PCs (dimensions) have been generated [86].

2.7.4 Computational steps of the multiple co-inertia analysis

The MCIA should be applied to the 57 cancer cell lines according simultaneously to both, gene and protein expression data of the first data sets from the HG-U133 plus 2.0 platform and proteome data. First, the two data frames has to be converted into a list and Aedín Culhane explained that one should check whether zero or low variant count rows are present in one of the used data sets [225]. In our case, the first data set of the HG-U133 plus 2.0 platform did not include zero or low variant count rows. The protein expression data contains 62 zero or low variant count rows and therefore this 62 rows were removed from this data set. Then the multiple co-inertia analysis was performed with the R-function *mcia()* [84,85] using the R-package *omicade4* [84,85].

The MCIA of the union of the top 50 DE genes, as well as proteins of the first data sets from the HG-U133 plus 2.0 platform and proteome data was performed in the same manner as the MCIA from all available genes and proteins. But in this case, only the top 50 DE genes and proteins were primarly selected from all the existing 36 tissue comparisons, and then the union of these genes or proteins was formed.

Finally, the most important genes at the positive and negative end of the first and second dimension were determined with the R-function *topVar()*.

2.8 Computational environment

The calculations and analysis of the gene and protein expression data from the NCI-60 panel were performed with the R version 3.3.2 [226]. The R-packages which were used in this study are shown in Table 4.

Table 4: Used R-packages for the calculations and analysis of the NCI-60 panel

R-package	**Package version**	**References**
affy	1.50.0	[135]
gcrma	2.44.0	[137]
hgu95a.db, hgu95b.db, hgu95c.db, hgu95d.db, hgu95e.db	3.2.3	[149], [150], [151], [152], [153]
hgu95acdf, hgu95bcdf, hgu95ccdf, hgu95dcdf, hgu95ecdf	2.18.0	[227], [228], [229], [230], [231]
hgu133a.db, hgu133b.db	3.2.3	[154], [155]
hgu133acdf, hgu133bcdf	2.18.0	[232], [233]
hgu133plus2.db	3.2.3	[156]
hgu133plus2cdf	2.18.0	[234]
openxlsx	4.0.0	[235]
Biobase	2.32.0	[236]
BiocGenerics	0.18.0	[237]
parallel	3.3.2	[238]
limma	3.28.6	[164]
DBI	0.5-1	[239]
AnnotationDbi	1.34.4	[240]
dynamicTreeCut	1.63_1	[241]
fastcluster	1.1.22	[242]
WGCNA	1.51	[243]
stats4	3.3.2	[244]
IRanges	2.6.0	[245]
S4Vectors	0.10.1	[246]
RSQLite	1.1-2	[247]
org.Hs.eg.db	3.3.0	[248]
annotate	1.50.1	[249]
Category	2.38.0	[195]
GOstats	2.38.1	[196]
ReactomePA	1.16.2	[201]
ade4	1.7-5	[219]
omicade4	1.12.0	[84]
XML	3.98-1.5	[250]
GO.db	3.3.0	[251]
BiocInstaller	1.22.3	[252]
sqldf	0.4-10	[253]

R-package	Package version	References
VennDiagram	1.6.17	[254]
graphics	3.3.2	[255]
data.table	1.10.0	[256]
scales	0.4.1	[257]
gplots_3.0.1	3.0.1	[175]
RColorBrewer_1.1-2	1.1-2	[176]
dplyr	0.7.1	[258]
rafalib	1.0.0	[259]
factoextra	1.0.4	[260]

3 Results

3.1 Sample Overlap between microarray and proteome data

The expression profile of the lung cancer cell line (LC:NCI_H23) is absent in the HG-U133 microarray platform. Furthermore the expression profile of the melanoma cancer cell line (ME:MDA_N) is missing in the HG-U133 plus 2.0 microarray platform as well as in the protein expression data set. Therefore, these two cancer cell lines were eliminated in all available microarray platforms. In the protein expression data set two other cancer cell lines are missing: ME:MDA_MB_435 and OV:NCI_ADR_RES, but just the cancer cell line OV:NCI_ADR_RES was removed from all four microarray platforms. Additionally there are cancer cell lines in the protein expression data set which are not covered by the microarray platforms: BR:MCF7ADR and BR:MDAMB435. It turned out that the breast cancer cell line, called BR:MDAMB435, actually corresponds to the melanoma cancer cell line ME:MDA_MB_435 (Meng C. personal communication). Thus, the name of the breast cancer cell line BR:MDAMB435 was changed to the melanoma cancer cell line ME:MDA_MB_435 in the protein expression data set and BR:MCF7ADR was deleted from this data set. Therefore, the number of cancer cell lines used in the analysis was reduced to 57 samples (cancer cell lines).

3.2 Comparison of annotations between original and re-annotated data table of the NCI-60 data set

The data sets downloaded from Cellminer are referred to original tables in the following text: they contain several duplicated probe IDs with same values. These duplicate rows were removed from these original data tables (Table 5 and Table 6). An overview of the different platforms and their probe IDs is shown in Table 5 and Table 6. The annotation of the generated data tables from the raw Affymetrix files of the four different microarray platforms were merged to the original Affymetrix data tables from Cellminer in order to check changes in the annotation (regarding probe ID – gene symbol pairs). In the next step the generated data sets were filtered. No suitable annotation R-package in Bioconductor exists for the Agilent 4x44K microarray and therefore, the original data table of the Agilent platform from Cellminer was compared to the annotation file of Agilent 4x44K downloaded from the webpage of Gene Expression Omnibus [157] (Table 6 and Table 8). Through the analysis we identified a few changes from probe ID – gene symbol pairs of the re-annotated generated tables to the probe ID – gene symbol pairs of the original tables from Cellminer (Table 7 and Table 8).

Table 5: Overview of the three different Affymetrix platforms

Platform	Total probe IDs in original data table	Total probe IDs in original data table without duplicated probe sets	Duplicated probe IDs in original data table	Duplicated control probe sets in original data table	Overlap of probe sets between different chips per platform
HG-U95	48,714	48,020	625	69	0
HG-U133	45,375	43,503	1,851	21	100
HG-U133 pus 2.0	54,703	52,691	2,012	0	0

Platform	Total probe IDs in re-annotated data table	Same probe IDs between original and re-annotated data tables	Probe IDs which just exist in re-annotated unfiltered data table
HG-U95	62,907	48,020	14,887
HG-U133	44,860	43,503	1,357
HG-U133 pus 2.0	54,675	52,691	1,984

Table 6: Overview of the Agilent platform

Platform	Total probe IDs in original data table	Total probe IDs in original data table without duplicated probe sets	Duplicated probe IDs in original data table	Duplicated control probe sets in original data table	Overlap of probe sets between different chips per platform
Agilent 4x44K	32,327	32,315	12	0	0

Platform	Total probe IDs in Annotation file	Same probe IDs between original data table and Annotation file	Probe IDs which just exist in Annotation file	Duplicated probe sets in downloaded Annotation file for the Re-annotation of Agilent data
Agilent 4x44K	41,094	32,315	8,779	4,126

Table 7: Changes in Affymetrix data between probe ID – gene symbol pairs of the re-annotated data tables and probe ID – gene symbol pairs of the original tables downloaded from Cellminer

Platform	Probe IDs where gene annotations remain same in re-annotated data table	Probe IDs where gene annotations are different in re-annotated data table	Probe IDs without gene annotation in re-annotated data table compared to original data table	Probe IDs present just in re-annotated data table only with gene annotation	Probe IDs present just in re-annotated data table only without gene annotation
HG-U95	22,732	6,009	19,279	12,136	2,751
HG-U133	30,295	4,527	8,681	669	688
HG-U133 plus 2.0	34,582	6,778	11,331	986	998

Table 8: Changes in Agilent data between probe ID – gene symbol pairs of the downloaded Annotation file and probe ID – gene symbol pairs of the original tables downloaded from Cellminer

Platform	Probe IDs from original data table where gene annotations are same in Annotation file	Probe IDs where gene annotations are different in Annotation file	Probe IDs without gene annotations in Annotation file	Probe IDs which just exist in Annotation file with gene annotation	Probe IDs which exist just in Annotation file without gene annotation
Agilent 4x44K	19,926	4,569	7,820	5,861	2,918

Data filtering is not really necessary for the performance of the different analysis methods, nevertheless, the microarray data were filtered in order to remove probes that do not belong to an official HUGO gene symbol, as well as eliminate control probes indicated by the "AFFX" term for Affymetrix (Table 9). This data filtering will later simplify the data interpretation. Furthermore, the number of probes that map to a single gene symbol, multiple (more) gene symbols, valid HUGO gene symbols was checked. Additionally unique gene symbols which map to multiple genes symbols were determined (Table 10).

Table 9: Filtering table

Platform	Control probes removed ("AFFX")	Probe IDs removed which do not map to an official HUGO gene symbol
HG-U95	90	19,274
HG-U133	21	8,670
HG-U133 plus 2.0	15	11,326
Agilent 4x44K	–	7,820
Protein expression	–	397

Table 10: Probe IDs mapping to gene symbols after filtering

Platform	Single mapping to gene symbol	Mappings to multiple gene symbols	Unique gene symbols from mappings to multiple gene symbols	Mappings to valid HUGO gene symbols
HG-U95	6,906	21,822	7,086	13,992
HG-U133	8,968	25,844	8,934	17,902
HG-U133 plus 2.0	9,591	31,759	10,608	20,199
Agilent 4x44K	11,584	12,911	5,613	17,197
Protein expression	6,544	1,172	543	7,087

In this project we generated three different data sets from the four available microarray platforms. In the first data set we excluded all control probes, as well as all probes which did not map to an official HUGO gene symbol. The second dataset contemplated probes that map to the same gene symbols (no unique mapping). In this case the probe with the highest average value across the used cell lines was selected. All probes which map to the same gene symbols were excluded in the third data set. The second and third data set represented the relation to the gene level. All three different kinds of data sets are summarized in Table 11.

Table 11: Summary of data sets

Platform	Cancer cell lines (samples)	First data set	Second data set	Third data set
HG-U95 (genes)	57	28,728	13,992	6,906
HG-U133 (genes)	57	34,812	17,902	8,968
HG-U133 plus 2.0 (genes)	57	41,350	20,199	9,591
Agilent 4x44K (genes)	57	24,495	17,197	11,584
Protein expression (proteins)	57	7,716	7,087	6,544

Compared to the gene expression data three different types of data sets of the protein expression were generated. Then the matching international protein index (IPI) identifiers, downloaded from the webpage of the Munich University of Technology [144], were assigned to the already processed and normalized label free-quantification protein expression data. Furthermore the corresponding HUGO gene symbols, downloaded from the webpage of the Munich University of Technology [144], were added to the IPI identifiers from the label free-quantification protein expression data. No further filtering and normalization steps were performed on the used data. We also investigated probes that map to a single, multiple (more) or valid HUGO gene symbol(s) (Table 10). The generation of the three different data sets was identical to the transciptome data sets (Table 11).

3.3 Overlap of genes (gene symbols) between different platforms

The Venn diagram shows common genes and proteins of the NCI-60 data set between the five processed platforms (HG-U95, HG-U133, HG-U133 plus 2.0, Agilent and proteome set). The overlap between the available platforms was based on gene symbol level

(Figure 3 and Figure 4). The Venn diagram considers only the unique gene symbols and therefore identical Venn diagrams were obtained for the first and second data sets (Figure 3).

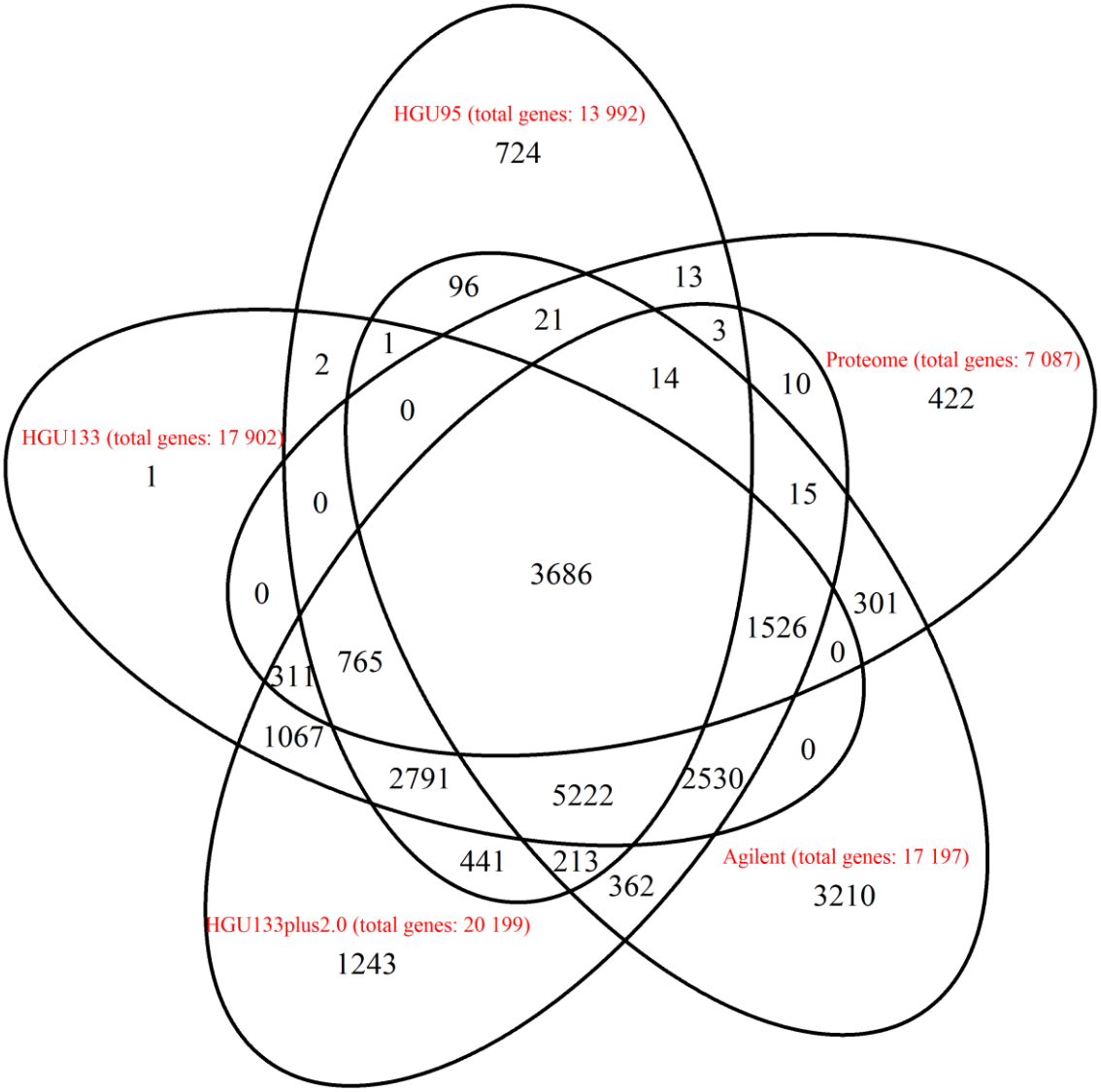


Figure 3: Venn diagram of the first and second data set which represents the overlapping genes and proteins of the NCI-60 data set from five preprocessed platforms. The total number of genes is 24,990. The Venn diagram considers only the unique gene symbols and therefore identical Venn diagrams were obtained for the first and second data sets.

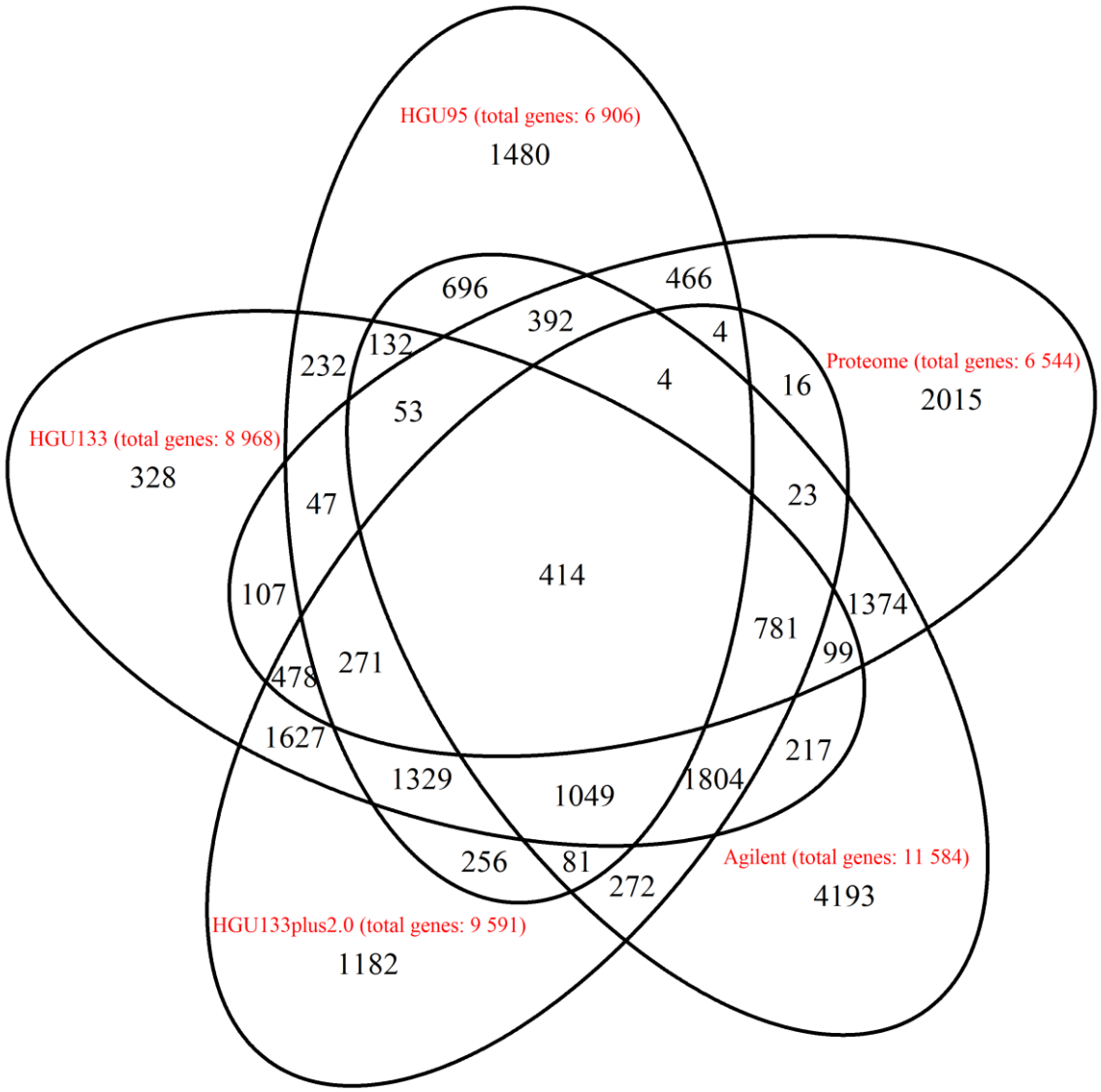


Figure 4: Venn diagram of the third data set which represents the overlapping genes and proteins of the NCI-60 data set from five preprocessed platforms. The total number of genes is 21,422.

3.4 Comparison of intensity differences from the three Affymetrix platforms

The histograms of the differences of expression data between the original and re-analyzed data tables from the three Affymetrix platforms (HG-U95, HG-U133 and HG-U133 plus 2.0) are shown in Figure 5. First the vector of differences of the expression values between the original and re-analyzed data tables from the three platforms over 57 cancer cell lines was generated. Further on these differences in expression data were represented by histograms (Figure 5).

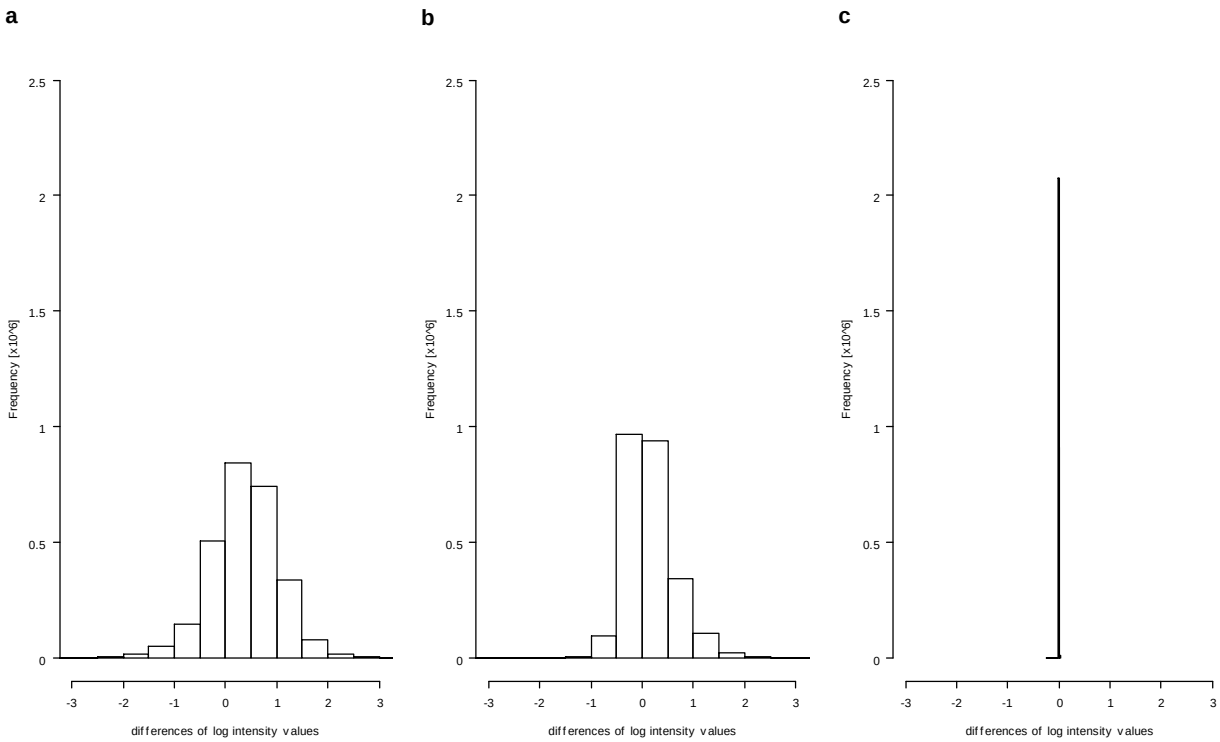


Figure 5: Histograms of differences of logarithmized intensity values (expression data) over 57 cancer cell lines between the original and re-analyzed data tables from the HG-U95 (a), HG-U133 (b) and platform HG-U133 plus 2.0 (c) platforms. Number of differences from expression data: 2,737,140 (a), 2,479,671 (b) and 3,003,387 (c) values.

3.5 Comparison of three different data sets from the HG-U133 plus 2.0 platform and proteome data

Table 12 represents an overview of all three different data set from the HG-U133 plus 2.0 platform. This data table shows and confirms that the first data set of the HG-U133 plus 2.0 platform contained almost all probe IDs or rather gene symbols, which also contained in the second and third data sets of the HG-U133 plus 2.0 platform. Therefore, from now on, only the first data set of the HG-U133 plus 2.0 platform will was analyzed, since compared to the other two data sets, it clearly contained the most information.

Table 12: Overview of three different data sets from HG-U133 plus 2.0 platform

Tissue comparison data	Length of probe IDs from first data set	Length of probe IDs from second data set	Length of probe IDs from third data set	Same probe IDs between first and second data set	Same probe IDs between first and third data set	Probe IDs which just exist in first data set compared to second data set	Probe IDs which just exist in first data set compared to third data set
BR_CNS	9	1	1	1	1	8	8
BR_CO	76	36	15	36	15	40	61
BR_LC	0	0	0	0	0	0	0
BR_LE	266	116	37	116	37	150	229
BR_ME	550	180	67	180	67	370	483
BR_OV	0	0	0	0	0	0	0
BR_PR	15	5	3	5	3	10	12
BR_RE	16	7	2	7	2	9	14
CNS_CO	2,742	1,241	502	1,241	502	1,501	2,240
CNS_LC	91	37	9	37	9	54	82
CNS_LE	2,348	966	306	966	306	1,382	2,042
CNS_ME	1,569	689	254	689	254	880	1,315
CNS_OV	258	118	44	118	44	140	214
CNS_PR	371	189	88	184	86	187	285
CNS_RE	438	163	57	163	57	275	381
CO_LC	303	126	40	126	40	177	263
CO_LE	1,079	483	174	483	174	596	905
CO_ME	3,149	1,361	530	1,361	530	1,788	2,619
CO_OV	68	33	8	33	8	35	60
CO_PR	120	60	28	60	28	60	92
CO_RE	2,467	1,053	388	1,053	388	1,414	2,079
LC_LE	1,395	557	184	557	184	838	1,211
LC_ME	974	428	170	428	170	546	804
LC_OV	0	0	0	0	0	0	0
LC_PR	25	11	4	11	4	14	21
LC_RE	8	1	0	1	0	7	8
LE_ME	3,210	1,252	429	1,252	429	1,958	2,781
LE_OV	817	333	102	333	102	484	715
LE_PR	330	165	72	165	72	165	258
LE_RE	3,439	1,401	468	1,401	468	2,038	2,971
ME_OV	1,239	524	208	524	208	715	1,031
ME_PR	646	303	128	303	128	343	518
ME_RE	2,875	1,278	471	1,278	471	1,597	2,404
OV_PR	27	10	6	10	6	17	21
OV_RE	19	6	2	6	2	13	17
PR_RE	163	69	34	69	34	94	129

Table 13 represents an overview of all three different data sets from the proteome data. This data table shows and confirms that the first data set of the proteome data contained almost all probe IDs or rather gene symbols, which are also contained in the second and third data sets of the proteome data. Therefore, from now on, only the first data set of the proteome data was analyzed, since compared to the other two data sets, it clearly contained the most information.

Table 13: Overview of three different data sets from proteome data

Tissue comparison data	Length of probe IDs from first data set	Length of probe IDs from second data set	Length of probe IDs from third data set	same probe IDs between first and second data set	same probe IDs between first and third data set	Probe IDs which just exist in first data set compared to second data set	Probe IDs which just exist in first data set compared to third data set
BR_CNS	7	7	5	7	5	0	2
BR_CO	3	3	3	3	3	0	0
BR_LC	0	0	0	0	0	0	0
BR_LE	24	23	22	23	22	1	2
BR_ME	1	0	0	0	0	1	1
BR_OV	1	1	1	1	1	0	0
BR_PR	69	66	64	66	64	3	5
BR_RE	3	3	3	3	3	0	0
CNS_CO	98	91	81	91	81	7	17
CNS_LC	1	1	1	1	1	0	0
CNS_LE	189	177	160	176	159	13	30
CNS_ME	21	19	17	19	17	2	4
CNS_OV	5	4	4	4	4	1	1
CNS_PR	83	80	76	80	76	3	7
CNS_RE	15	15	15	15	15	0	0
CO_LC	14	13	12	13	12	1	2
CO_LE	89	86	80	85	80	4	9
CO_ME	85	82	78	82	78	3	7
CO_OV	1	1	1	1	1	0	0
CO_PR	45	44	44	44	44	1	1
CO_RE	47	42	39	42	39	5	8
LC_LE	105	94	87	94	87	11	18
LC_ME	2	2	2	2	2	0	0
LC_OV	0	0	0	0	0	0	0
LC_PR	28	28	27	28	27	0	1
LC_RE	2	2	2	2	2	0	0
LE_ME	139	131	121	130	120	9	19
LE_OV	97	93	82	92	82	5	15
LE_PR	142	135	127	135	127	7	15
LE_RE	134	119	102	119	102	15	32
ME_OV	1	3	0	1	0	0	1
ME_PR	39	38	36	38	36	1	3
ME_RE	37	36	34	35	33	2	4
OV_PR	32	30	30	30	30	2	2
OV_RE	1	1	1	1	1	0	0
PR_RE	31	31	30	31	30	0	1

3.6 Differential expression analysis

Table 14 represents the tissue comparison data (36 sample comparisons), the DE genes/proteins (over-expressed and under-expressed) based on the p-value ($\leq$ 0.05), as well as the positive/negative log2-fold change of the gene and protein expression data of the first data set from the HG-U133 plus 2.0 platform and the proteome data. The tissue comparison data (36 sample comparisons), the DE genes/proteins (over-expressed and under-expressed) based on the p-value ($\leq$ 0.05), as well as the positive/negative log2-fold change of the gene and protein

expression data of the second and third data set from the HG-U133 plus 2.0 platform and the proteome data are shown in the Appendix (Chapter 6; Supplementary Tables 1 – 2). In the Appendix (Chapter 6) there are also detailed tables, which additionally have a column with the total number of genes/proteins, with the number of DE genes/proteins (over-expressed and under-expressed) and with the number of the unchanged genes/proteins with respect to the p-value (> 0.05) for three different cut-offs of the absolute $\log_2$-fold change values (1, 1.4 and 2) (Supplementary Tables 3 – 8).

Table 14: DEA of the gene expression data of the first data set from the HG-U133 plus 2.0 platform with absolute $\log_2$-fold change 1 in the upper triangular matrix and DEA of the protein expression data of the first data set from the proteome data with absolute $\log_2$-fold change 1 in the lower triangular matrix

	BR	CNS	CO	LU	LE	ME	OV	PR	RE
BR		2; 7	46;30	0; 0	188; 78	217; 333	0; 0	1; 14	4; 12
CNS	2; 5		1,477; 1265	49; 42	1,545; 803	853; 716	148; 110	124; 247	193; 245
CO	1; 2	55; 43		145; 158	758; 321	1,540; 1,609	23;45	14; 106	1,078; 1,389
LU	0; 0	1; 0	11; 3		843; 552	411; 563	0; 0	2; 23	0; 8
LE	16; 8	150; 39	68; 21	77; 28		1,347; 1,863	240; 557	41; 289	1,455; 1,984
ME	1; 0	15; 6	56; 29	2; 0	33; 106		656; 583	173; 473	1,369; 1,506
OV	0; 1	3; 2	0; 1	0; 0	30; 67	0; 1		7; 20	3; 16
PR	4; 65	20; 63	13; 32	2; 26	9; 133	3; 36	5; 27		128; 35
RE	0; 3	6; 9	21; 26	1; 1	35; 99	17; 20	0; 1	27; 4	

3.7 Correlation of expression profiles

The 9x9 heatmaps of the mean Pearson correlation coefficients according to gene and protein expression data between the nine different tissues and the 57x57 heatmaps of the mean Pearson correlation coefficients according to gene and protein expression data between the 57 cancer cell lines of the first data set from the HG-U133 plus 2.0 platform and proteome data are represented in Figures 6 – 9. The heatmaps implement also dendrograms in the rows which show the hierarchical clustering of the different samples. In the Appendix (Chapter 6) the 9x9 heatmaps of the mean Pearson correlation coefficients according to gene and protein expression data between the nine different tissues and the 57x57 heatmaps of the mean Pearson correlation coefficients according to gene and protein expression data between the 57 cancer cell lines of the second and third data set from the HG-U133 plus 2.0 platform and proteome data are shown in Supplementary Figures 1 – 8.

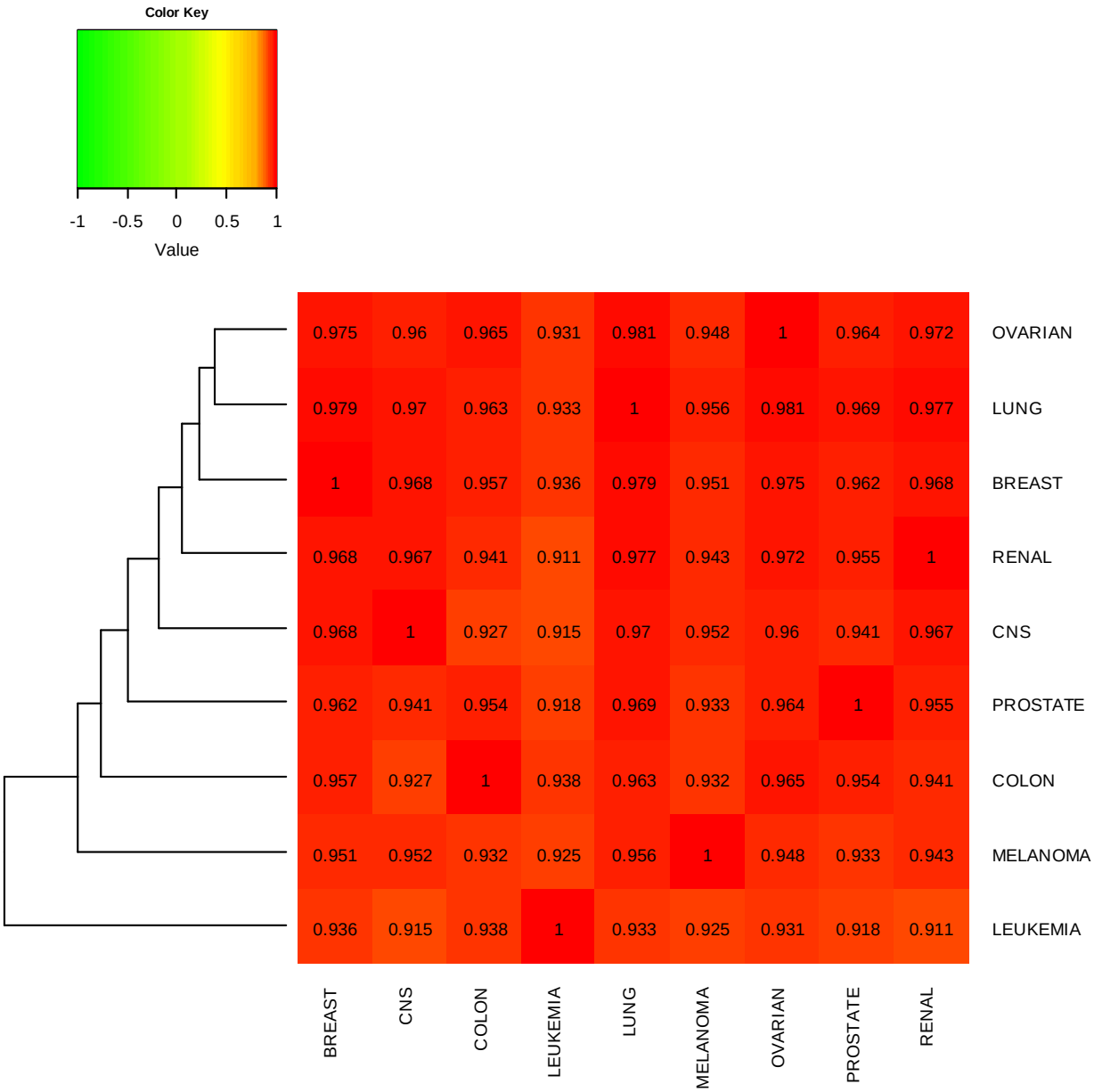


Figure 6: Heatmap of the mean Pearson correlation coefficients of the first data set of gene expression data of 41,350 mRNA probe sets between the nine different tissues where the rows of the heatmap are arranged by the hierarchical clustering according to the average linkage method of 9 tissue origins in the NCI-60 data.

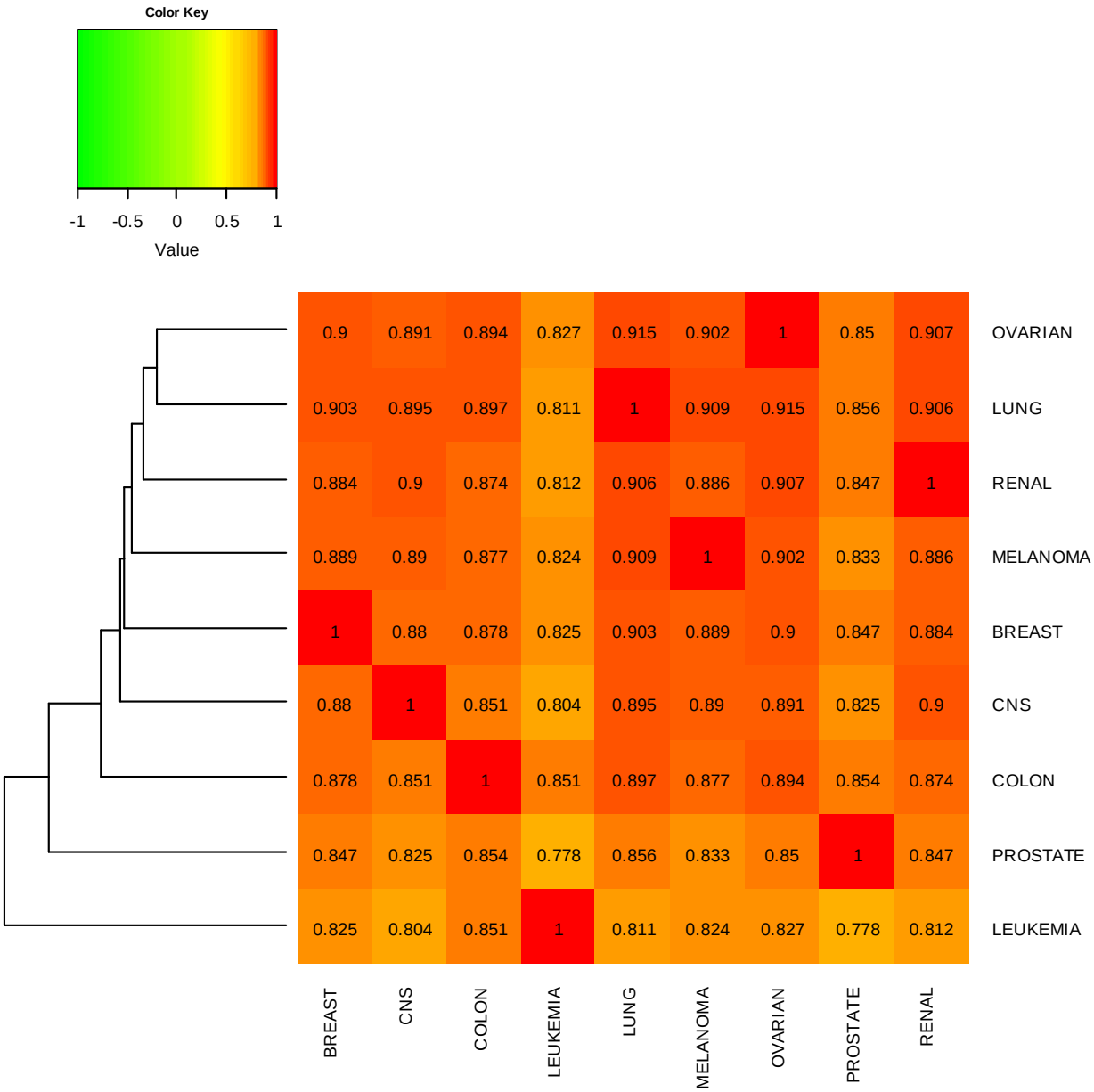


Figure 7: Heatmap of the mean Pearson correlation coefficients of the first data set of protein expression data of 7,716 proteins where between the nine different tissues the rows of the heatmap are arranged by the hierarchical clustering according to the average linkage method of 9 tissue origins in the NCI-60 data.

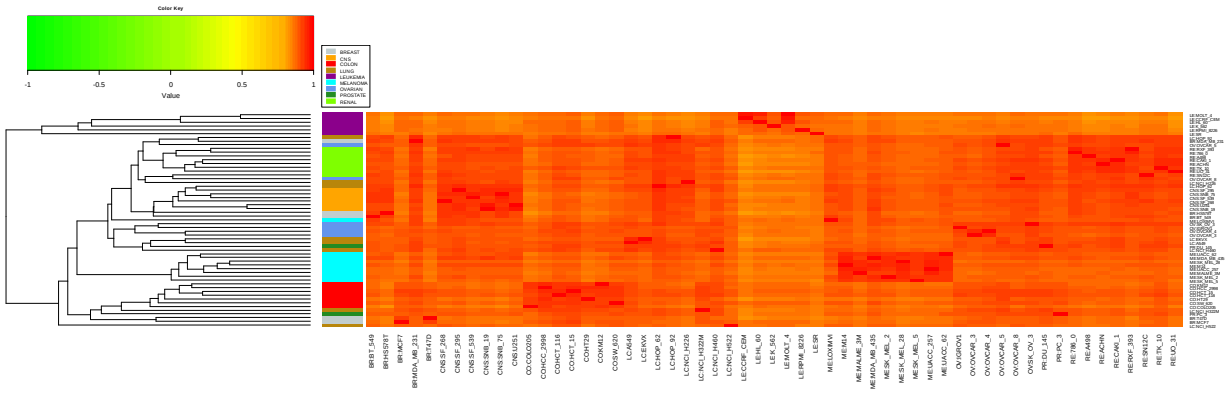


Figure 8: Heatmap of the mean Pearson correlation coefficients of the first data set of gene expression data of 41,350 mRNA probe sets between the 57 different cancer cell lines where the rows of the heatmap are arranged by the hierarchical clustering according to the average linkage method of the 57 cancer cell lines in the NCI-60 data.

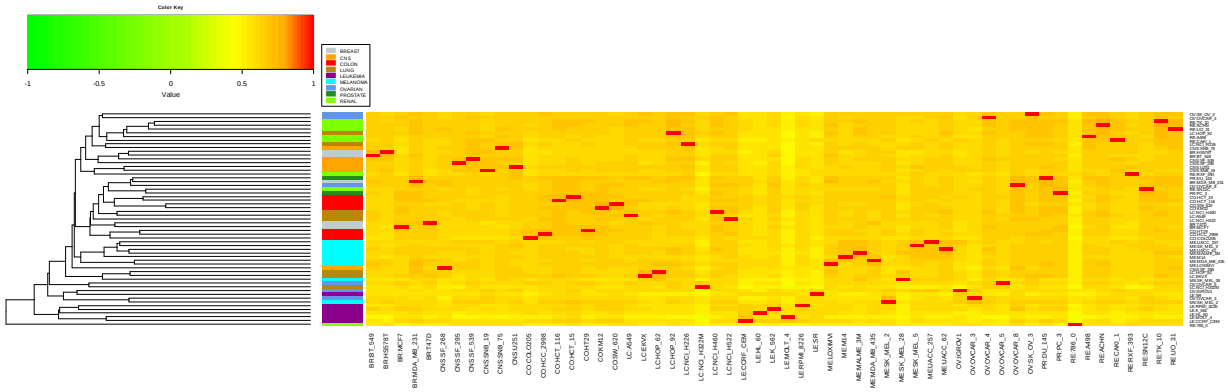


Figure 9: Heatmap of the mean Pearson correlation coefficients of the first data set of protein expression data of 7,716 proteins between the 57 different cancer cell lines where the rows of the heatmap are arranged by the hierarchical clustering according to the average linkage method of the 57 cancer cell lines in the NCI-60 data.

The dendrograms of the 57 cancer cell lines of the first data set according to gene and protein expression data generated using hierarchical clustering are represented in Figures 10 – 11. Therefore the samples, which belong to the same tissue origins, are plotted in the same colors (Figures 10– 11). In the Appendix (Chapter 6) the dendrograms of the 57 cancer cell lines of the second and third data set according to gene and protein expression data generated using hierarchical clustering are represented in Supplementary Figures 9 – 12.

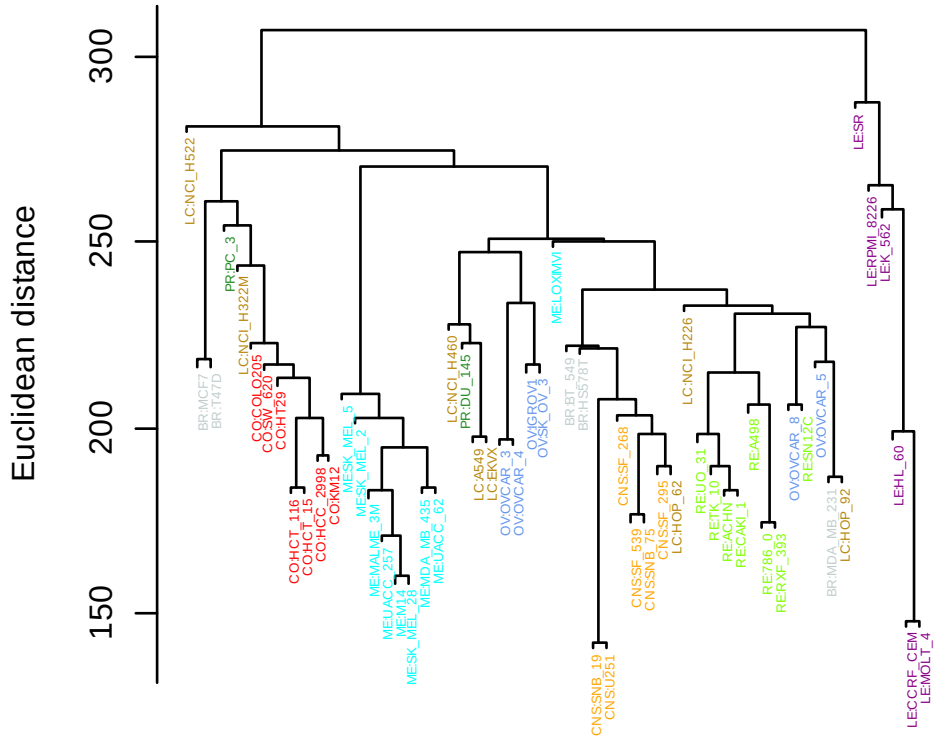


Figure 10: Dendrogram of 57 cancer cell lines of the first data set of gene expression data of 41,350 mRNA probe sets where the samples of the dendrogram are arranged by the hierarchical clustering according to the average linkage method of the 57 cancer cell lines in the NCI-60.

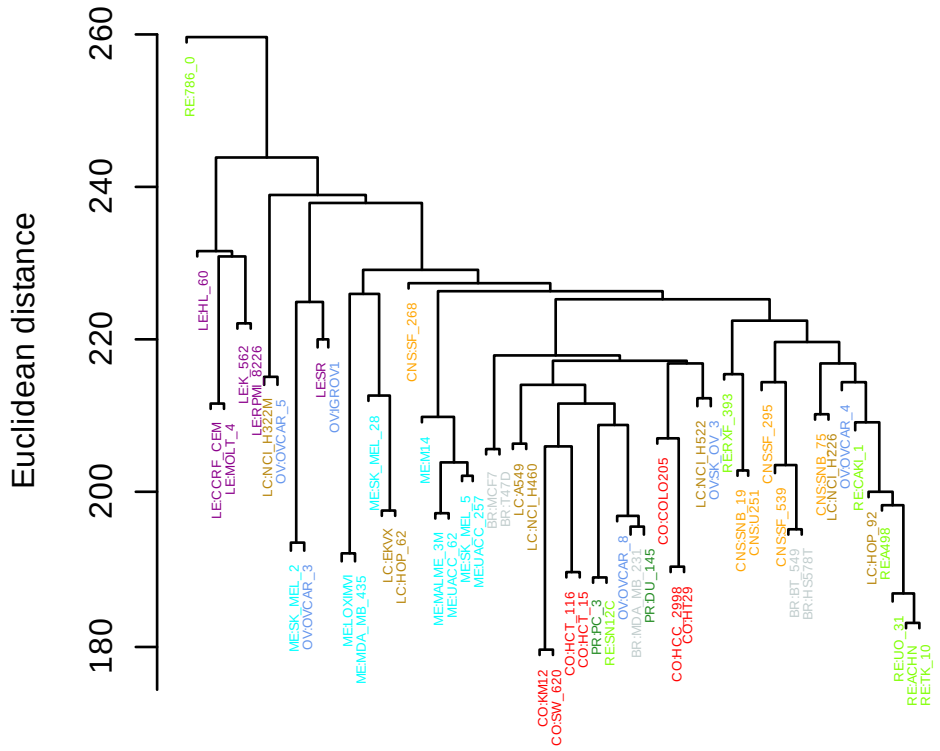


Figure 11: Dendrogram 57 cancer cell lines of the first data set of protein expression data of 7,716 proteins where the samples of the dendrogram are arranged by the hierarchical clustering according to the average linkage method of the 57 cancer cell lines in the NCI-60 data.

3.8 Gene set enrichment analysis of gene ontology categories and Reactome pathways

The results of GSEA of DE genes/proteins of the different tissue comparisons of each GO category (biological process (BP), molecular function (MF) and cellular component (CC)) and Reactome pathways of the first data set from the HG-U133 plus 2.0 platform and proteome data are represented in Tables 15 – 16. The results of GSEA of DE genes/proteins of the different tissue comparisons of each GO category (biological process (BP), molecular function (MF) and cellular component (CC)) and Reactome pathways of the second and third data set from the HG-U133 plus 2.0 platform and proteome data are represented in the Appendix (Chapter 6; Supplementary Tables 9 – 12).

Table 15: Results of GSEA of the first data set from HG-U133 plus 2.0 platform

Tissue comparison data	Number of DE genes	Number of members from BP GO category	Number of members from MF GO category	Number of members from CC GO category	Number of members from Reactome pathways
BR_CNS	9	15	13	4	3
BR_CO	76	336	43	38	0
BR_LE	0	0	0	0	0
BR_LC	266	795	106	94	14
BR_ME	550	1,051	125	90	0
BR_OV	0	0	0	0	0
BR_PR	15	36	12	1	16
BR_RE	16	124	17	29	0
CNS_CO	2,742	975	140	201	21
CNS_LE	91	376	67	68	1
CNS_LC	2,348	1,340	149	191	48
CNS_ME	1,569	1,154	146	171	14
CNS_OV	258	632	112	112	9
CNS_PR	371	389	107	107	10
CNS_RE	438	613	114	130	0
CO_LE	303	458	89	51	0
CO_LC	1,079	1,228	190	140	67
CO_ME	3,149	1,496	224	183	5
CO_OV	68	407	53	56	0
CO_PR	120	448	69	46	0
CO_RE	2,467	857	140	139	9
LC_LE	1,395	1,198	126	151	67
LC_ME	974	1,380	194	154	0
LC_OV	0	0	0	0	0
LC_PR	25	104	7	6	0
LC_RE	8	201	12	13	5
LE_ME	3,210	1,249	210	177	19
LE_OV	817	1,437	153	181	42
LE_PR	330	1,045	97	105	20
LE_RE	3,439	1,411	198	241	90
ME_OV	1,239	1,343	187	160	3
ME_PR	646	639	102	80	0
ME_RE	2,875	1,477	192	225	10
OV_PR	27	115	18	22	0
OV_RE	19	171	11	20	0
PR_RE	163	417	68	40	0

Table 16: Results of GSEA of the first data set from proteome data

Tissue comparison data	Number of DE genes	Removed proteins which do not have corresponding Affymetrix or ENTREZ ID	Number of members from BP GO category	Number of members from MF GO category	Number of members from CC GO category	Number of members from Reactome pathways
BR_CNS	7	0	0	60	16	7
BR_CO	3	1	1	59	13	17
BR_LE	0	0	0	0	0	0
BR_LC	24	3	3	442	51	73
BR_ME	1	0	NA[3]	NA[5]	NA[5]	Warning[4]
BR_OV	1	0	NA[5]	NA[5]	NA[5]	Warning[6]
BR_PR	69	4	4	251	79	76
BR_RE	3	0	0	84	7	17
CNS_CO	98	7	7	925	109	133
CNS_LE	1	0	NA[5]	NA[5]	NA[5]	Warning[6]
CNS_LC	189	14	14	1,069	165	222
CNS_ME	21	1	1	293	45	54
CNS_OV	5	0	0	198	22	12
CNS_PR	83	3	3	253	62	58
CNS_RE	15	0	0	156	27	40
CO_LE	14	2	2	245	36	27
CO_LC	89	2	2	711	97	108
CO_ME	85	4	4	523	86	73
CO_OV	1	1	1	0	0	0
CO_PR	45	3	3	269	54	48
CO_RE	47	4	4	497	67	63
LC_LE	105	7	7	549	130	161
LC_ME	2	0	0	80	8	27
LC_OV	0	0	0	0	0	0
LC_PR	28	0	0	264	44	40
LC_RE	2	1	1	NA[5]	NA[5]	Warning[6]
LE_ME	139	9	9	607	124	144
LE_OV	97	2	2	599	98	114
LE_PR	142	7	7	481	96	118
LE_RE	134	11	11	1,100	120	198
ME_OV	1	0	NA[5]	NA[5]	NA[5]	Warning[6]
ME_PR	39	1	1	305	48	42
ME_RE	37	0	0	558	86	71
OV_PR	32	2	2	137	30	44
OV_RE	1	0	NA[5]	NA[5]	NA[5]	Warning[6]
PR_RE	31	0	0	167	41	51

[3] Error when using only 1 de gene in GSEA for GO categories – Error: "Error in getGoToEntrezMap_db(p): genes being tested do not have corresponding GO terms"

[4] Warning in GSEA for Reactome pathways – Warning: "No gene can be mapped …. → return NULL…"

3.9 Principal component analysis of all available genes or proteins from the HG-U133 plus 2.0 platform and protein expression data

Table 17 shows the proportion of variances of the first five PCs of all available genes or proteins of the first data set from the HG-U133 plus 2.0 platform, as well as the proteome data (protein expression data).

Table 17: Proportion of variances explained by the first five PCs of all available genes or proteins and of the union of the top 50 DE genes or proteins (in %)

Data sets of different data	Probe IDs/gene symbols	Cancer cell lines	PC1	PC2	PC3	PC4	PC5
HG-U133 plus 2.0 first data set from all samples	41,350	57	13.6	10.3	7.2	3.9	3.6
Protein expression first data set from all samples	7,716	57	6.3	5.8	4.5	4	3.7
HG-U133 plus 2.0 first data set from DE genes/proteins	738	57	24.9	19.6	9.7	5.3	3.3
Protein expression first data set from all genes/proteins	485	57	13	9.3	6.2	4.7	3.8

The first and second principal components (PC1, PC2) of the 57 cancer cell lines according to gene and protein expression data of the first data set from the HG-U133 plus 2.0 platform and proteome data are represented in Figure 12. The x-axis denotes the PC 1 which shows how much variance is explained by the PC1. The y-axis indicates the PC2 which represents how much variance is explained by the PC2. The 57 samples (cancer cell lines), which belong to the same tissue origins, are plotted in the same colors (Figure 12).

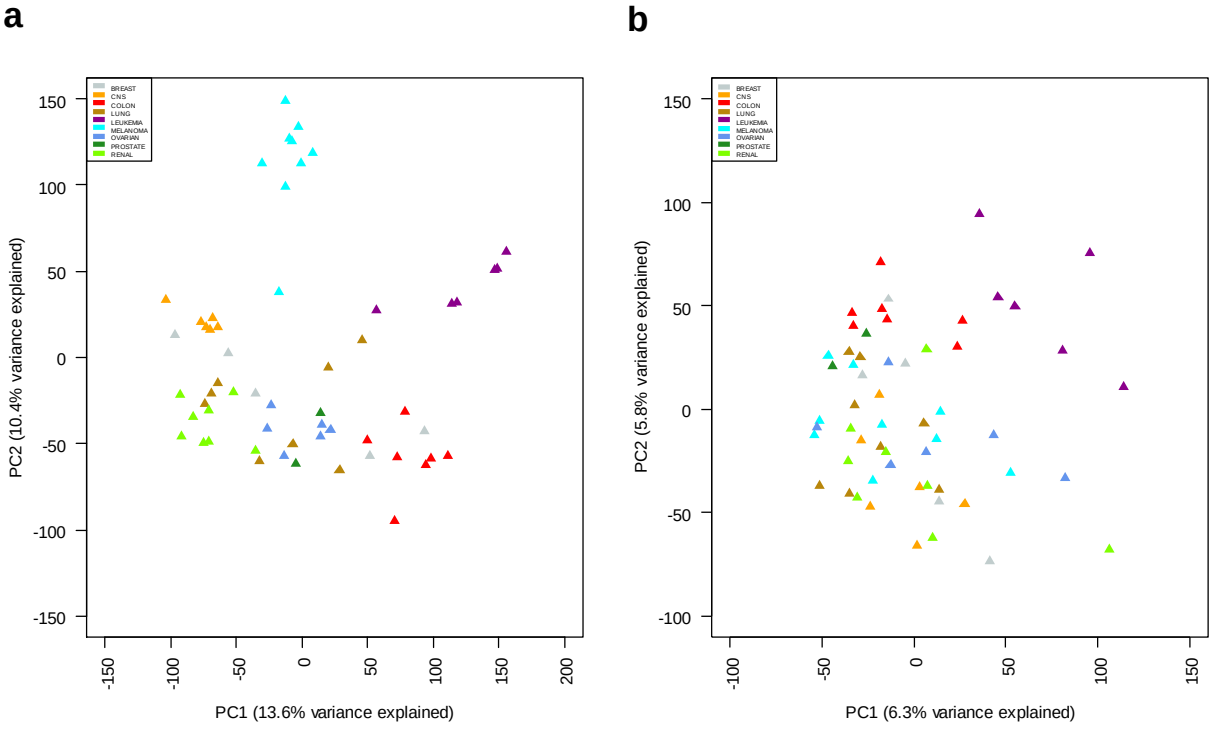


Figure 12: PCA of PC1 and PC2 of the 57 samples according to gene (a) and protein expression (b) data of the first data set from the HG-U133plus 2.0 platform and proteome data.

3.10 Principal component analysis of the union of the top 50 differentially expressed genes or proteins from the HG-U133 plus 2.0 platform and protein expression data

Table 17 shows the proportion of variances of the first five PCs of the union of the top 50 DE genes or proteins of the first data set from the HG-U133 plus 2.0 platform, as well as the proteome data (protein expression data).

The PC1 and PC2 of the 57 cancer cell lines according to the union of the top 50 DE genes or proteins of the first data set from the HG-U133 plus 2.0 platform and proteome data are represented in Figure 13. The x-axis denotes the PC1 which shows how much variance is explained by the PC1. The y-axis indicates the PC2 which represents how much variance is explained by the PC2. The 57 samples (cancer cell lines), which belong to the same tissue origins, are plotted in the same colors (Figure 13).

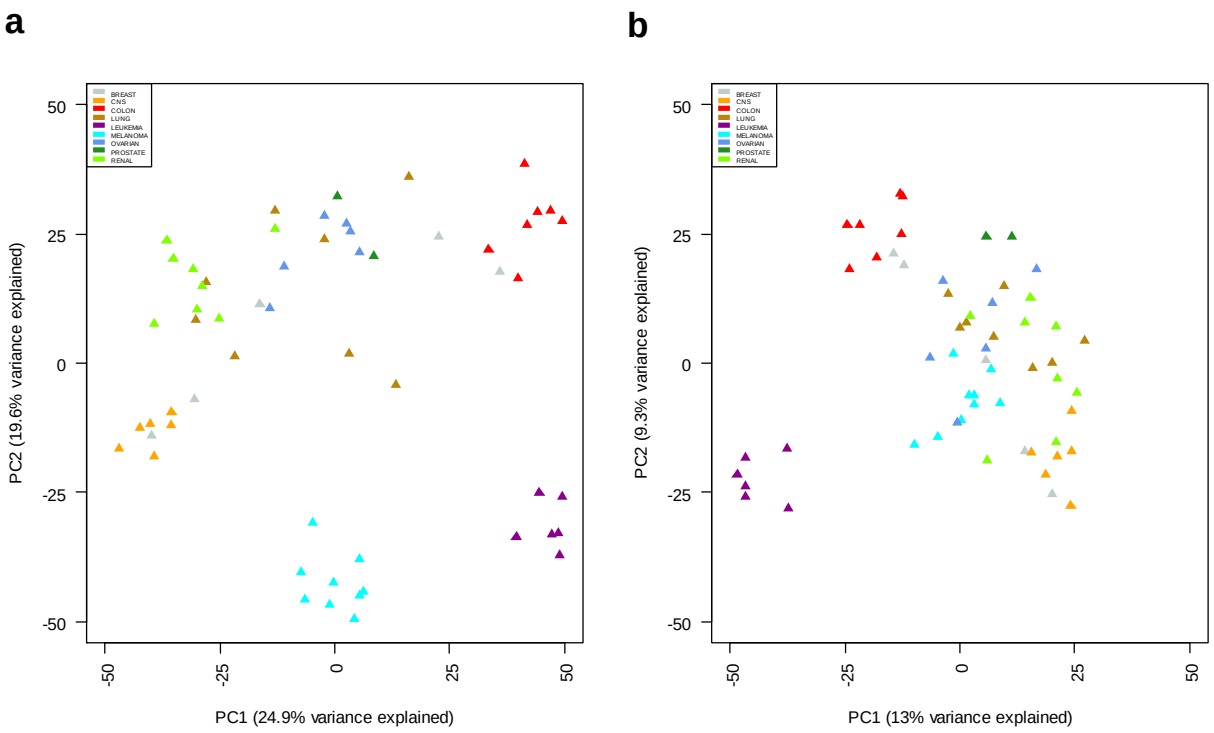


Figure 13: PCA of PC1 and PC2 of the 57 samples according to the union of the top 50 DE genes (a; total of 738 genes) or proteins (b; total of 485 proteins) of the first data set from the HG-U133plus 2.0 platform and proteome data.

3.11 Correspondence analysis of all available genes or proteins from the HG-U133 plus 2.0 platform and protein expression data

Table 18 shows the proportion of variances of the first five dimensions of all available genes or proteins of the first data set from the HG-U133 plus 2.0 platform and proteome data (protein

expression data), as well as the total inertias of these data sets which correspond to the sum of the eigenvalues.

Table 18: Proportion of variances explained by the first five dimensions of all available genes or proteins and of the union of the top 50 DE genes or proteins (in %)

Data sets of different data	Probe IDs/gene symbols	Cancer cell lines	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5	Total inertia
HG-U133 plus 2.0 first data set from all samples	41,350	57	12	10.2	7.1	3.9	3.5	~0.04
Protein expression first data set from all samples	7,716	57	4.7	3.8	3.6	3.1	3	~1.10
HG-U133 plus 2.0 first data set from DE genes/proteins	738	57	23.1	17.5	11.3	5.7	3.3	~0.15
Protein expression first data set from DE genes/proteins	485	57	13.2	8.4	6.4	4.3	3.8	~1.11

The first and second dimension of the 57 cancer cell lines according to gene and protein expression data of the first data set from the HG-U133 plus 2.0 platform and proteome data are represented in Figure 14. The x-axis denotes the first dimension which shows how much variance is explained by the first dimension. The y-axis indicates the second dimension which represents how much variance is explained by the second dimension. The 57 samples (cancer cell lines), which belong to the same tissue origins, are plotted in the same colors (Figure 14).

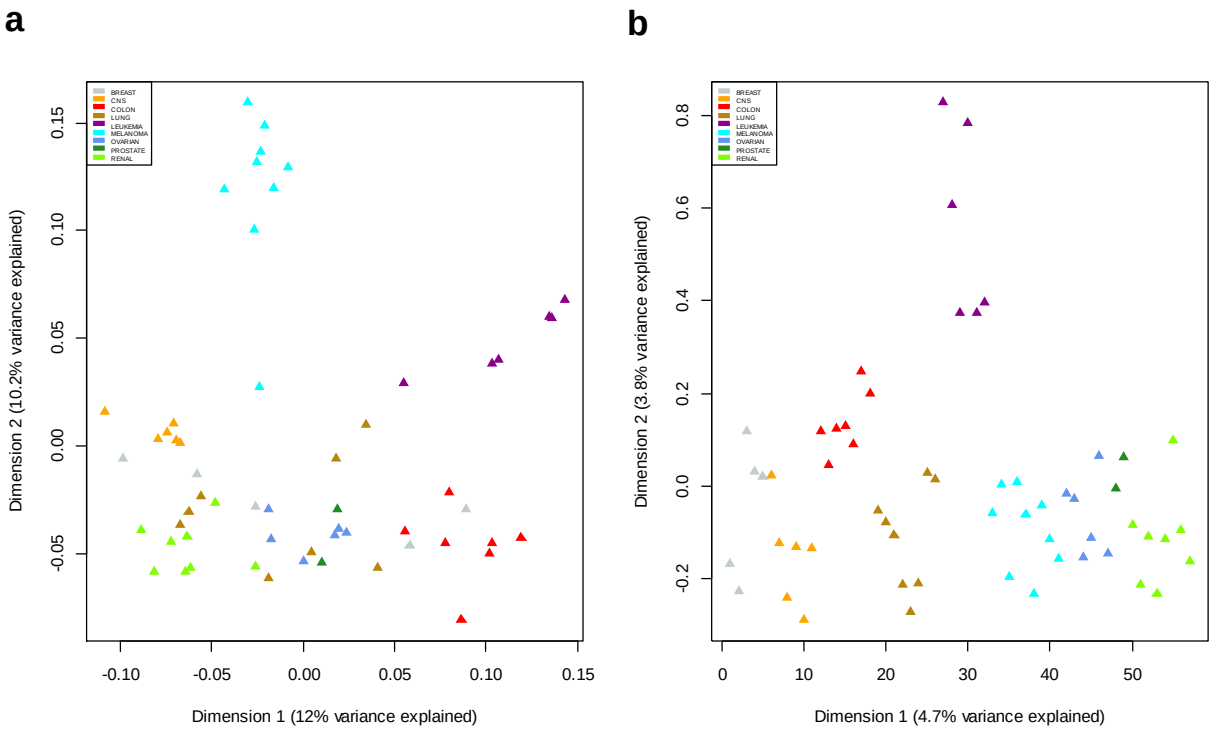


Figure 14: CA of first and second dimension of the 57 samples according to gene (a) and protein (b) expression data of the first data set from the HG-U133plus 2.0 platform and proteome data.

3.12 Correspondence analysis of the union of the top 50 differentially expressed genes or proteins from the HG-U133 plus 2.0 platform and protein expression data

Table 18 shows the proportion of variances of the first five dimensions of the union of the top 50 DE genes or proteins of the first data set from the HG-U133 plus 2.0 platform and proteome data (protein expression data), as well as the total inertias of these data sets.

The first and second dimension of the 57 cancer cell lines according to the union of the top 50 DE genes or proteins of the first data set from the HG-U133 plus 2.0 platform and proteome data are represented in Figure 15. The x-axis denotes the first dimension which shows how much variance is explained by the dimension. The y-axis indicates the second dimension which represents how much variance is explained by the second dimension. The 57 samples (cancer cell lines), which belong to the same tissue origins, are plotted in the same colors (Figure 15).

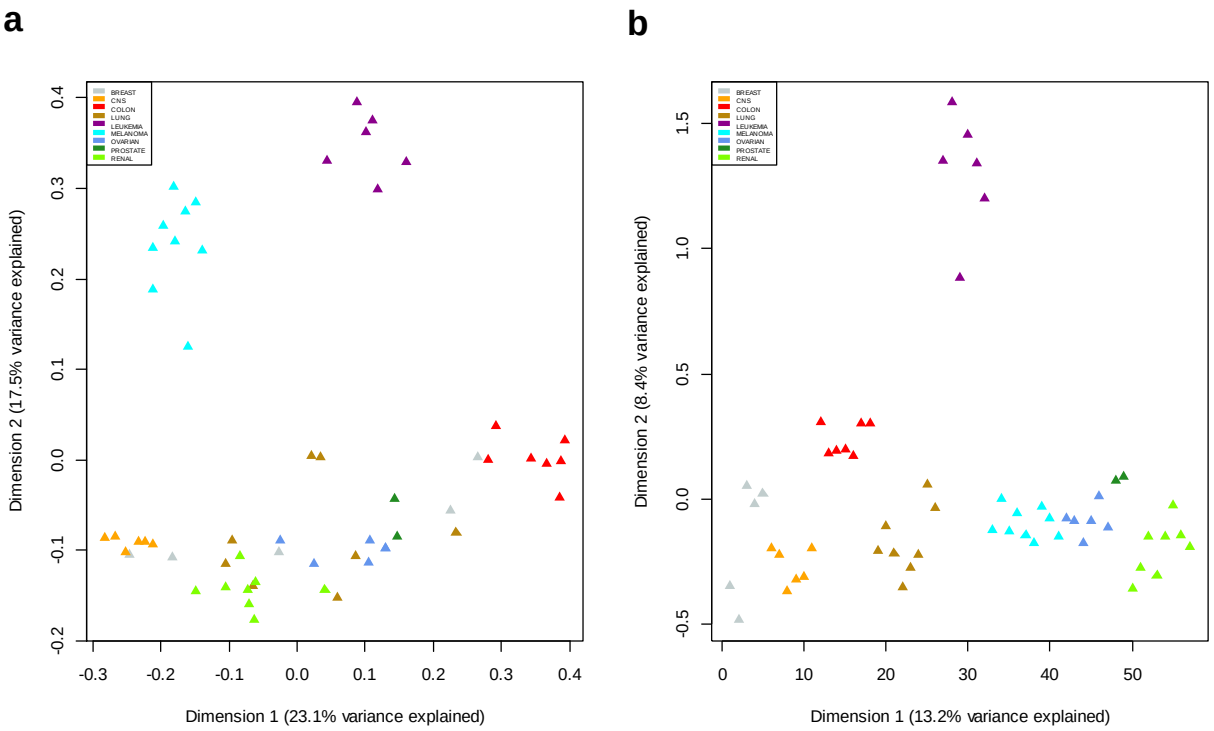


Figure 15: CA of first and second dimension of the 57 samples according to the union of the top 50 DE genes (a; total of 738 genes) or proteins (b; total of 485 proteins) of the first data set from the HG-U133plus 2.0 platform and proteome data.

3.13 Non-symmetrical correspondence analysis of all available genes or proteins from the HG-U133 plus 2.0 platform and protein expression data

Table 19 shows the proportion of variances of the first five dimensions of all available genes or proteins of the first data set from the HG-U133 plus 2.0 platform, as well as the proteome data (protein expression data).

Table 19: Proportion of variances explained by the first five dimensions of all available genes or proteins and of the top 50 DE genes or proteins (in %)

Data sets of different data	Probe IDs/gene symbols	Cancer cell lines	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5
HG-U133 plus 2.0 first data set from all samples	41,350	57	12	10.2	7.1	3.9	3.5
Protein expression first data set from all samples	7,716	57	4.6	3.9	3.6	3.1	3
HG-U133 plus 2.0 first data set from DE genes/proteins	738	57	23.5	17	10.1	6.1	3.3
Protein expression first data set from DE genes/proteins	485	57	11	7.7	6.9	5.3	4.1

The first and second dimension of the 57 cancer cell lines according to gene and protein expression data of the first data set from the HG-U133 plus 2.0 platform and proteome data are represented in Figure 16. The x-axis denotes the first dimension which shows how much variance is explained by the first dimension. The y-axis indicates the second dimension which represents how much variance is explained by the second dimension. The 57 samples (cancer cell lines), which belong to the same tissue origins, are plotted in the same colors (Figure 16).

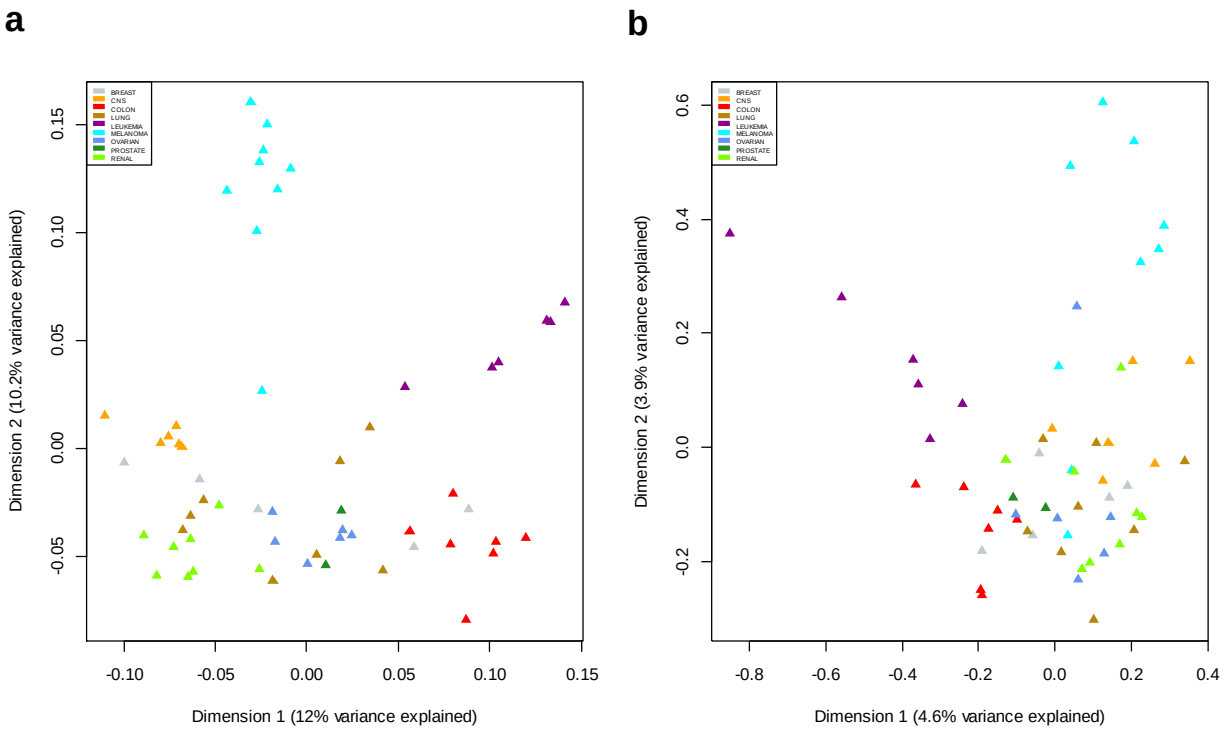


Figure 16: NSCA of first and second dimension of the 57 samples according to gene (a) and protein (b) expression data of the first data set from the HG-U133plus 2.0 platform and proteome data.

3.14 Non-symmetrical correspondence analysis of the union of the top 50 differentially expressed genes or proteins from the HG-U133 plus 2.0 platform and protein expression data

Table 19 shows the proportion of variances of the first five dimensions of the union of the top 50 DE genes or proteins of the first data set from the HG-U133 plus 2.0 platform, as well as the proteome data (protein expression data).

The first and second dimension of the 57 cancer cell lines according to the union of the top 50 DE genes or proteins of the first data set from the HG-U133 plus 2.0 platform and proteome data are represented in Figure 17. The x-axis denotes the first dimension which shows how much variance is explained by the dimension. The y-axis indicates the second dimension which represents how much variance is explained by the second dimension. The 57 samples (cancer cell lines), which belong to the same tissue origins, are plotted in the same colors (Figure 17).

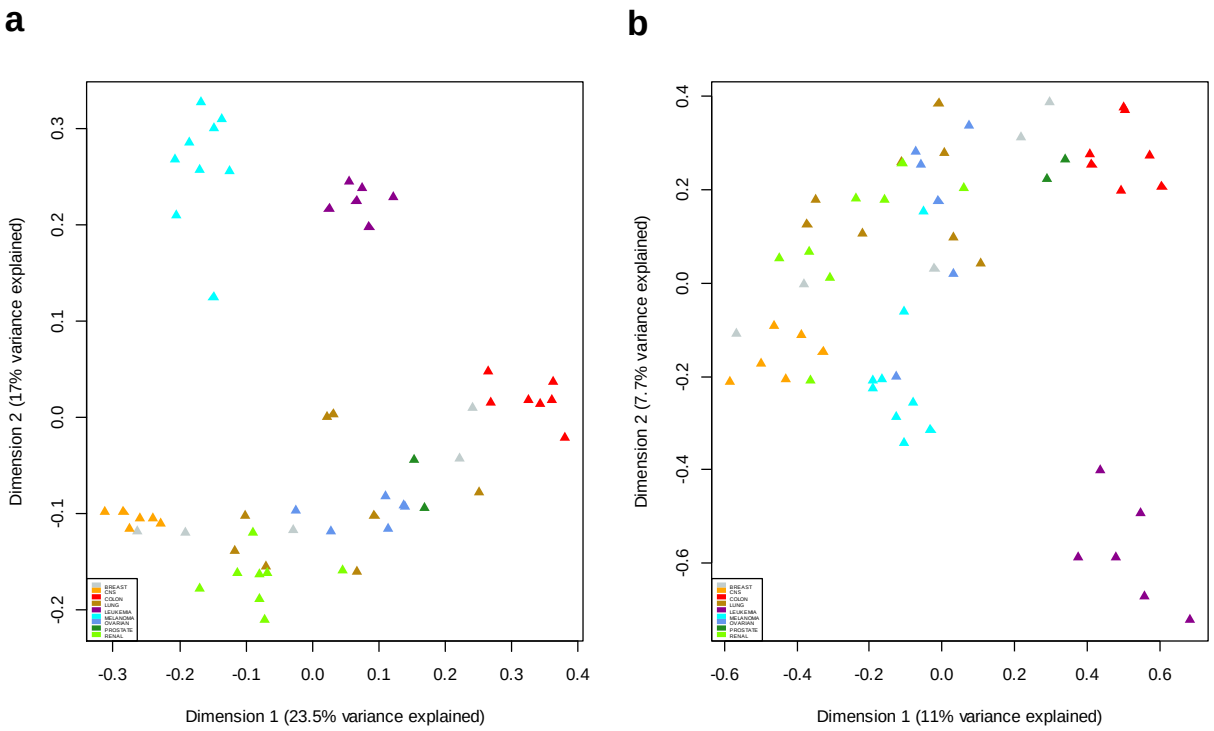


Figure 17: NSCA of first and second dimension of the 57 samples according to the union of the top 50 DE genes (a; total of 738 genes) or proteins (b; total of 485 proteins) of the first data set from the HG-U133plus 2.0 platform and proteome data.

3.15 Multiple co-inertia analysis of all available genes or proteins from the HG-U133 plus 2.0 platform and protein expression data

Table 20 shows the proportion of variances of the first five dimensions of all available genes and proteins of the first data set from the HG-U133 plus 2.0 platform, as well as the proteome data (protein expression data).

Table 20: Proportion of variances explained by the first five dimensions of all available genes and proteins and of the union of the top 50 DE genes and proteins (in %)

Data sets of different data	Probe IDs/gene symbols of gene expression	Probe IDs/gene symbols of protein expression	Cancer cell lines	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5
HG-U133 plus 2.0 and protein expression of first data set	41,350	7,654	57	16.1	13.6	10.6	6.3	5.9
HG-U133 plus 2.0 and protein expression of DE genes and proteins of first data set	738	485	57	33.2	24	16.8	9.9	7.4

The first and second dimension of the 57 cancer cell lines according to both, gene and protein expression data of the first data set from the HG-U133 plus 2.0 platform and proteome data are represented in Figure 18. The x-axis denotes the first dimension which shows how much variance is explained by the first dimension. The y-axis indicates the second dimension which represents how much variance is explained by the second dimension. The 57 samples (cancer cell lines) of each platform, which belong to the same tissue origins, are plotted in the same colors and the two different platforms, HG-U133 plus 2.0 platform and proteome data, are represented by different symbols in the figure (Figure 18).

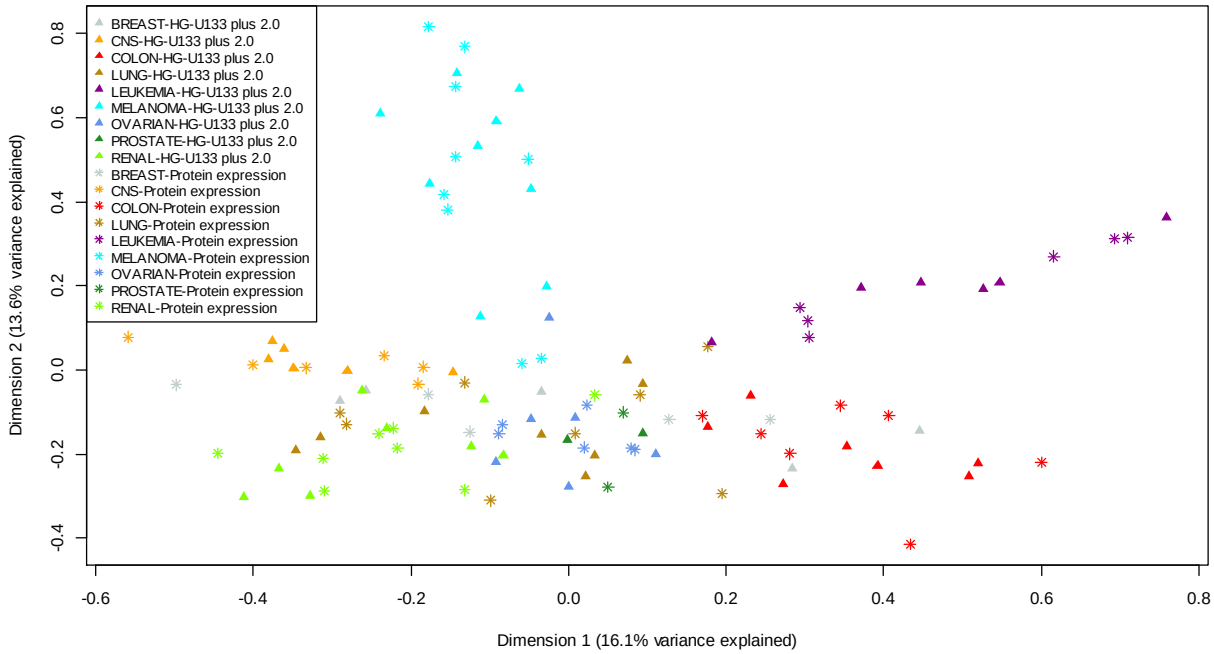


Figure 18: MCIA of first and second dimension of the 57 samples according to gene and protein expression data of the first data set from the HG-U133plus 2.0 platform and proteome data.

3.16 Multiple co-inertia analysis of the union of the top 50 differentially expressed genes or proteins from the HG-U133 plus 2.0 platform and protein expression data

Table 20 shows the proportion of variances of the first five dimensions of the union of the top 50 DE genes and proteins of the first data set from the HG-U133 plus 2.0 platform, as well as the proteome data (protein expression data).

The first and second dimension of the 57 cancer cell lines according to the union of the top 50 DE genes and proteins of the first data set from the HG-U133 plus 2.0 platform and proteome data are represented in Figure 19. The x-axis denotes the first dimension which shows how much

variance is explained by the dimension. The y-axis indicates the second dimension which represents how much variance is explained by the second dimension. The 57 samples (cancer cell lines) of each platform, which belong to the same tissue origins, are plotted in the same colors and the two different platforms, HG-U133 plus 2.0 platform and proteome data, are represented by different symbols in the figure (Figure 19).

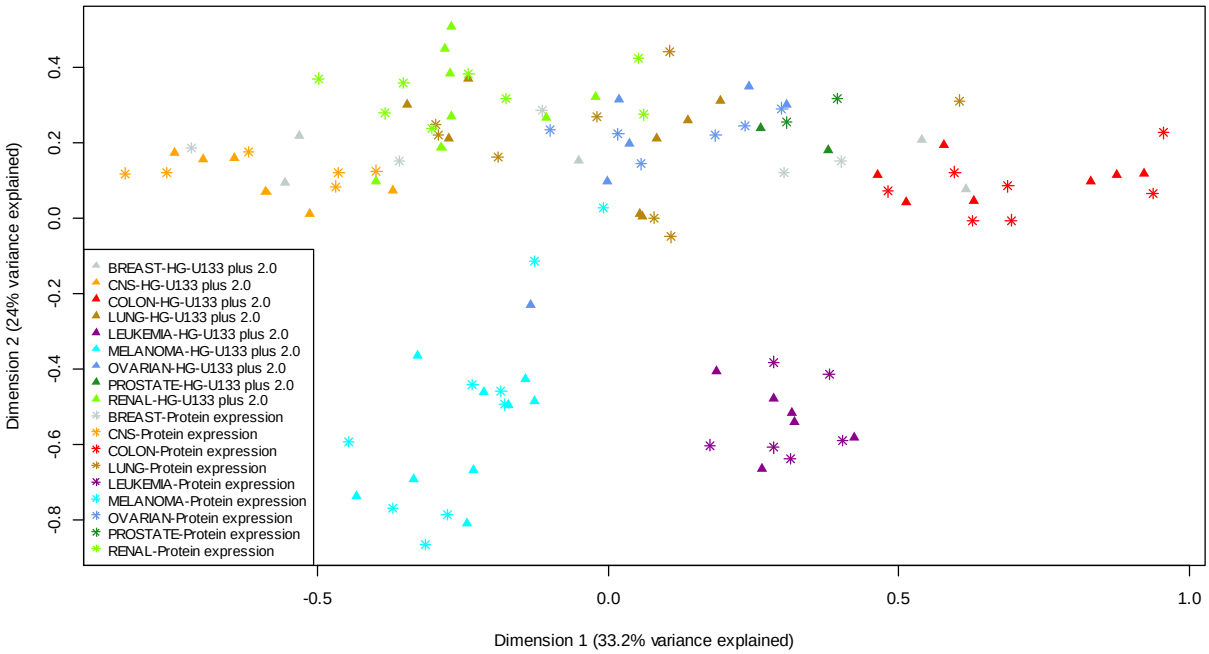


Figure 19: MCIA of first and second dimension of the 57 samples according to the union of the top 50 DE genes (total of 738 genes) and proteins (total of 485 proteins) of the first data set from the HG-U133plus 2.0 platform and proteome data.

3.17 Most influential genes on the positive and negative end of the first principal component/dimension and the second principal component/dimension

The most influential genes on the positive and negative end of PC1/D1 and PC2/D2 from the three sequential analysis methods from all available genes (PCA_{ALL}, CA_{ALL}, $NSCA_{ALL}$) based on gene expression and the integrative analysis method from all available genes and proteins ($MCIA_{ALL}$) are shown in Table 21. It can be clearly seen that the different methods highlighted similar genes on the ends of PC1/D1 and PC2/D2 (Table 21).

Table 21: Most influential genes on the positive and negative end of PC1/D1 and PC2/D2 from the three sequential analysis methods from all available genes (PCA_{ALL}, CA_{ALL}, $NSCA_{ALL}$) based on transcriptome data and the integrative analysis method from all available genes and proteins ($MCIA_{ALL}$)

	positive end PC1/D1 (colon and breast marker)	positive end PC1/D1 (epithelial marker)	positive end PC1/D1 (leukemia marker)	negative end PC1/D1 (mesenchymal marker)	positive end PC2/D2 (melanoma marker)
PCA_{ALL}	MYB, SYTL1, ESRP1, CORO1A	S100P, EPCAM; SLC27A2	HCLS1	SPARC, GJA1, FN1	MLANA, DCT, TYR, S100B
CA_{ALL}	MYB, CORO1A, RAB25	S100P, RAB25	PTPRC, HCLS1	SPARC, COL1A2, COL8A1, COL5A1, COL1A1	MLANA, DCT, TYR
$NSCA_{ALL}$	MYB, CORO1A, RAB25	S100P, RAB25	PTPRC, HCLS1	SPARC, COL1A2	MLANA, DCT, TYR
$MCIA_{ALL}$	MYB, CORO1A	S100P	PTPRC, HCLS1	SPARC, COL1A2	MLANA, DCT, TYR, S100B

The most influential genes on the positive and negative end of PC1/D1 and PC2/D2 from the three sequential analysis methods from the union of top 50 DE genes (PCA_{50DE}, CA_{50DE}, $NSCA_{50DE}$) based on gene expression and the integrative analysis method from the union of top 50 DE genes and proteins ($MCIA_{50DE}$) are shown in Table 22. It can be clearly seen that the different methods highlighted similar genes on the ends of PC1/D1 and PC2/D2 (Table 22).

Table 22: Most influential genes on the positive and negative end of PC1/D1 and PC2/D2 from the three sequential analysis methods from the union of top 50 DE genes (PCA_{50DE}, CA_{50DE}, $NSCA_{50DE}$) based on transcriptome data and the integrative analysis method from the union of top 50 DE genes and proteins ($MCIA_{50DE}$)

	positive end PC1/D1 (epithelial marker)	negative end PC1/D1 (mesenchymal marker)	positive end PC2/D2 (leukemia marker)	positive end PC2/D2 (melanoma marker)	negative end PC2/D2 (leukemia marker)	negative end PC2/D2 (melanoma marker)
PCA_{50DE}	CDH1, EPCAM, KRT19, SPINT2, SLC27A2	SPARC, GJA1, COL4A1	–	–	BCL2A1	MLANA, BCL2A1
CA_{50DE}	S100A14, EPCAM, CDH1	SPARC, COL1A2	HCLS1, BCL2A1, SOX10	MLANA, BCL2A1	–	–
$NSCA_{50DE}$	S100A14, EPCAM, CDH1	SPARC, COL1A2	HCLS1, BCL2A1	MLANA, BCL2A1	–	–
$MCIA_{50DE}$	CDH1, EPCAM, KRT19, SPINT2, SLC27A2	SPARC, COL1A2, COL5A1, COL6A2	–	–	HCLS1, PTPRC	MLANA, BCL2A1, SOX10, RAB38

4 Discussion

The goal of this study was to conduct a comparison between sequential and integrative methods for the analysis of omics data. In this thesis we demonstrated the power of the three different sequential analysis methods PCA, CA and NSCA, as well as the integrative analysis method MCIA, which had been applied to the transcriptome and proteome levels of the NCI-60 data. To the best to our knowledge, no direct comparison between sequential and integrative analysis methods had been reported so far. The comparison of these analysis methods confirmed that integrative analysis is more powerful than corresponding sequential analysis and yielded more information than a single data set alone. While sequential analysis methods focus only on single (individual) data sets, integrative analysis methods provide a deeper insight and comprehensive understanding of a biological system due to the integration of multi-level omics data (multiple layers of information).

4.1 Re-analysis and Re-annotation

The human genome sequence was first released in February 2001 [261], followed by the final sequence in April 2003 [262,263]. Since then, the sequence was updated regularly with the human reference genome GRCh38 being the latest release from the Genome Reference Consortium in December 2013 [264]. Such regular updates are also applied to GO [265]. Annotation packages in Bioconductor [148] are updated about every six months based on the information in these release databases. With the re-annotation process the original gene annotation will be updated and the information remains accurate, relevant and always up-to-date. It enables the determination of relationships between probes on various microarrays and genes by mapping probe IDs to gene annotations, such as HUGO gene symbols, or gene-level identifiers, such as Entrez IDs or GO IDs, which are available in the associated annotation packages in Bioconductor. An interesting aspect of this is that some annotation databases, which are among the most commonly used ones, integrate outdated gene annotations in pathway enrichment analysis [266]. Usually, the re-annotation using appropriate regularly updated annotation databases allows a more accurate and precise analysis of microarray data compared to the original gene annotations provided by the manufacturer. In addition, updates of GO IDs facilitate a better interpretation of data. Another important aspect is the result of continuous improvements in methodology which enable more accurate assigning with respect to the gene annotation.

In this study gene expression data from the following platforms, HG-U95, HG-U133, HG-U133 plus 2.0 and Agilent, were re-analyzed. This allowed the comparison of the re-annotated and re-analyzed data tables (called "updated data tables" in the following) to the already processed, normalized, analyzed and annotated original data tables, which can be downloaded directly from the Cellminer webpage (called "original data tables" in the further course). The re-analysis revealed that the gene expression values of the updated data table of the HG-U133 plus 2.0 platform showed the smallest deviations with regard to the gene expression levels of the original data table compared to the HG-U95 and HG-U133 platforms (Figure 5). Thus, the re-analysis of the Affymetrix HG-U133 plus 2.0 raw data resulted in similar values of expression levels as the analysis of the original data table from this platform. Gene annotation changed significantly between original and updated Affymetrix data tables for all three platforms, as well as the original Agilent data table and the Annotation file for the Agilent 4x44K platform. This was mainly due to the fact that in the updated data tables a large number of probe IDs had no gene annotation in contrast to the original tables (Tables 5 – 8). During the re-annotation of the HG-U133 plus 2.0 platform it was noted that this platform had the most recent scan/measurement date (Table 3).

The re-analysis and re-annotation results and the most recent scan/measurement date led to the decision that the data set of the HG-U133 plus 2.0 platform is the most appropriate for further analyzes.

4.2 Differential expression analysis and gene set enrichment analysis

The 57 cancer cell lines were grouped according to tissue origin in order to determine DE genes/proteins between the nine different tissue types. Melanoma cell lines exhibited the largest number of DE genes in contrast to the other tissue categories, whereas breast cancer cell lines showed the least number of DE genes compared to the other tissue groups (Table 14). Leukemia cell lines showed the most DE proteins compared to the other cancer types, whereas the cancer cell lines originating from breast tissue had the least number of DE proteins, closely followed by ovarian and lung cancer cells, in contrast to the other carcinomas (Table 14). The Pearson correlation coefficients of average gene and protein expression abundances demonstrated the levels of global variability and similarities within and between different tissue types of the NCI-60 data (Figures 6 – 7). Expression correlation and the number of DE genes/proteins showed an inverse relationship: if the Pearson correlation coefficient in the respective tissue group was relatively low, the number of DE genes/proteins had a relatively high value, and vice

versa (Table 14 and Figures 6 – 7). The Pearson correlation coefficients of protein expression data between the various tissue categories showed greater variability ($R \geq 0.778$; Figure 7) than the Pearson correlation of gene expression measurements ($R \geq 0.911$; Figure 6). It could be seen that none of the cancer cell lines of the same tissue origin based on gene expression were distinctly different from another tissue category (Figure 6), whereas Figure 7 showed that the leukemia cancer cell lines according to protein expression were clearly different from the other tissue types, closely followed by the prostate cancer cell lines. In summary, however, it could be observed that the Pearson correlation coefficients between the various tissue categories had high values, both based on gene and protein expression data. This indicated a high co-structure between the tissue types of different cancer cell lines.

Our findings demonstrated that gene and protein expression profiling could well differentiate between different cancer types of various tissue groups in most cases. Furthermore, the overrepresented GO annotations and pathways received from DE genes and proteins showed that the transcriptome, respectively proteome, reflected the main functions and pathways of the appropriate cancer tissue type. For example, GSEA of the breast cancer cell lines highlighted some significant GO terms regarding biological processes, such as “mRNA translation including translational termination”, “positive/negative regulation of immune response”, “cell adhesion”, “innate immune response”, “biological adhesion” and “innate immune response”, which were consistent with the findings in a previous study [267]. Furthermore, some significant enriched biological processes related to breast cancer cells had been found, such as “positive regulation of biological process”, “positive regulation of cellular process”, “cellular response to organic substance” and “positive regulation of transcription from RNA polymerase II promoter”, which confirmed the results of an earlier study [268].

4.3 Clustering analysis of mRNA and protein expression data

In general it could be observed that cancer cell lines derived from the same tissue origin usually converged in the same or closely related cluster, whereas cancer cell lines originating from different tissue categories showed greater variability and were spread across multiple clusters.

Clustering of mRNA expression data showed that cancer cell lines belonging to leukemia (six out of six; clearly separated from the other tissue groups) and melanoma (eight out of nine), were grouped into the same cluster which characterizes their tissue of origin (Figures 8 and 10). The majority of CNS, colon and renal cell lines of gene expression were grouped in clusters according to their tissue of origin (Figures 8 and 10). These findings confirmed the results of

other studies, where leukemia, CNS; colon and melanoma cancer cell lines achieved a clear separation by tissue origin [269–274]. Ovarian cell lines were inhomogenously clustered, two of the six samples, OV:OVCAR_5 and OV:OVCAR_8, did not belong to the same cluster (Figures 8 and 10). Lung, as well as prostate cancer cell lines were not clustered homogenously (Figures 8 and 10). Cancer cell lines originating from the breast tissues were widely distributed over several clusters. Previous studies showed that cancer cell lines originating from leukemia, melanoma, CNS, colon and renal tissues could be considered similar and had grouped together in a cluster in earlier expression profiles studies, whereas breast, lung and prostate cancer cell lines spread over multiple clusters [272,274,275]. The two hormone dependent estrogen receptor-positive (ER+) breast cancer cell lines, BR:MCF7 and BR:T47D, were arranged in a cluster compared to the other breast cancer cell lines [274,276]. The breast cancer cell lines can be divided into two subgroups: luminal, BR:MCF7 and BR:T47D, and basal, BR:HS578T, BR:BT_549 and BR:MDA_MB_231 [276]. On the one hand, the luminal cell lines are positive for the estrogen receptor [274]. The basal cell lines, on the other hand, are negative for hormone receptors and are characterized by basal cytokeratins [271]. Correlation within the tissue origin was higher than the correlations between tissues of various origins (Figures 6 and 8). This observation is consistent with the findings of other studies [269,272,274,275]. The high correlation among the two leukemia cancer cell lines, LE:CCRF_CEM and LE:MOLT_4, that are acute lymphocytic leukemias (ALL), was an obvious case for disease sub-typing inside a tissue type [269]. The other leukemia cancer cell lines, such as LE:HL_60 (acute promyelocytic leukemia), LE:K_562 (chronic myelogenous leukemia), LE:RPMI_822 (multiple myeloma) and LE:SR (large cell immunoblastic lymphoma), showed a lower correlation as the two above-mentioned leukemia cancer cells and no subgrouping. The melanoma cancer cell line, ME: LOXIMVI, tended to cluster with various tissue groups (Figures 6 and 8). This could also be observed in earlier studies [272,274,275]. This melanoma cancer cell line was recognized to be highly differentiated and amelanotic and therefore did not express characteristic melanocyte genes [277]. It had been known that this melanoma cancer cell type produces no melanin and lacks other typical markers of melanoma cells [277,278].

Based on protein expression, the majority of cancer cell lines (CNS, colon, leukemia and melanoma) were grouped closely in clusters according to their tissue of origin (Figures 9 and 11). Several subclusters were formed: leukemia (five out of six), CNS (four out of six), colon (four out of seven) and melanoma (five out of nine) which for the most part were grouped based on their tissue origin (Figure 11). In contrast, cancer cell lines originating from breast, lung, ovarian, prostate and renal tissues showed inconsistent clustering

(Figures 9 and 11). These cancer cell lines were widely distributed and grouped in multiple clusters. This pointed out that their protein expression patterns are quite heterogeneous. Similar observations were made in a previous study [141]. Clustering of luminal and basal breast cancer cell lines, the two ALL cell lines, and the melanoma cell line ME: LOXIMVI were consistent with the results in the cluster analysis of the gene expression data.

In general, it was observed that clustering based on protein expression had a greater level of variation and inconsistency compared to clustering based on gene expression.

4.4 Sequential analysis methods

Sequential analysis methods, also called DRTs, enable the representation of high-dimensional omics data, such as measurements (samples, arrays) and features (genes, mRNAs, microRNAs, proteins, metabolites), in a lower dimensional space. Visualization of the results of these analysis methods allows the identification of the most influential features within data sets and therefore facilitates the interpretation of organisms on a global scale. A limitation of these methods is that they only allow the analysis of single individual data sets. In this study, we investigated PCA [88,100–103], CA [104–107] NSCA [106,108,109]. These analysis methods were already used in order to analyze gene and protein expression data [89,101–103,105].

4.4.1 PCA from all available genes and proteins and from the union of top 50 DE genes and proteins

Both for PCA from all available genes and proteins (PCA_{ALL}; Figure 12), as well as from the union of top 50 DE genes and proteins (PCA_{DE50}; Figure 13), it could be seen that the first two PCs in gene expression level explained a larger proportion of variance (PCA_{ALL}: PC1: 13.6% and PC2: 10.3% and (PCA_{DE50}: PC1: 24.9% and PC2: 19.6%) compared to these in protein expression patterns (PCA_{ALL}: PC1: 6.3% and PC2: 5.8% and PCA_{DE50}: PC1: 13% and PC2: 9.3%).

Cancer cell lines originating from the same tissue origin based on gene expression were close to each other, converged in clusters and were therefore distinct from other organs. Based on gene expression the majority of cell lines from melanoma (eight out of nine), leukemia (five out of six), CNS (five out of six), colon (six out of seven) and renal (eight out of nine) tissues were grouped in clusters according to their tissue of origin. The findings of the PCA results regarding all breast cancer cell lines, the two leukemia cancer cell lines and one melanoma cell line (ME:

LOXIMVI, which lacks melanogenesis) based on gene expression data were consistent with the results in the cluster analysis. Breast, lung and ovarian tissues were widely distributed which meant that their gene expression patterns appeared most heterogeneous. The two prostate cancer cell lines (PR:DU_145 and PR:PC_3) did not cluster close together. PR:DU_145, closely projected to the ovarian cancer cell lines, showed similar results as hierarchical clustering and a previous finding [279]. PCA also revealed that estrogen receptor negative (ER–) breast cancer cell lines (BR:HS578T and BR:BT_549) were grouped with CNS malignancies and that ER+ breast cancer lines (BR:MCF7 and BR:T47D) were closely clustered to colon cancer lines, consistent with results of a previous study [141].

The PCA_{DE50} based on transcriptome data resulted in more pronounced clustering of cancer cell lines originating from same tissue compared to the PCA_{ALL} based on gene expression (Figure 12). Again, the separation of cancer cell lines according to PCA_{DE50} based on proteome was less pronounced in the protein expression profiles compared to the mRNA profile, which was consistent with the results of PCA_{ALL}.

On PC1 based on gene expression of PCA_{ALL} the mesenchymal features, such as secreted protein acidic and cysteine rich (SPARC), gap junction protein alpha 1 (GJA1) [280] and fibronectin 1 (FN1) [281], were separated from colon and breast cancer cell lines associated with MYB proto-oncogene, transcription factor (MYB) [282], synaptotagmin like 1 (SYTL1) [283], also known as secretory leukocyte protease inhibitor (SLPI), epithelial splicing regulatory protein 1 (ESRP1) [284,285], also known as RNA-binding motif protein 35A (RBM35A), and coronin 1A (CORO1A) [286], as well as from epithelial markers, such as S100 calcium binding protein P (S100P) [287,288], epithelial cell adhesion molecule (EPCAM), which is a cell adhesion molecule and occurs in most healthy epithelial cells, and solute carrier family 27 member 2 (SLC27A2) and also from leukemia-related genes, such as hematopoietic cell-specific Lyn substrate 1 (HCLS1). Melanoma-related genes based on gene expression of PCA_{ALL}, such as melan-A (MLANA) [289], dopachrome tautomerase (DCT) [289], tyrosinase (TYR) [289,290] and S100 calcium-binding protein B (S100B) [289], were projected on the positive side of PC2. Thus melanoma cancer cells were separated from the others on PC2.

PCA_{DE50} selected the most influential genes associated with positive values of PC1, including SLC27A2, Epithelial-cadherin (CDH1), EPCAM, serine peptidase inhibitor, Kunitz type 2 (SPINT2) and keratin 19 (KRT19), which can be regarded as epithelial markers. This analysis led to similar results regarding to epithelial markers as in [112]. The mesenchymal markers, such as SPARC and GJA1, and collagen markers, including collagen type IV alpha 1 chain

(COL4A1), associated with negative values were projected on PC1. The epithelial and mesenchymal features were separated along the horizontal axis. PCA_{DE50} showed that the leukemia and melanoma cancer cell lines were separated on PC2 from the other carcinomas. Therefore, the melanocyte-specific genes, including MLANA and BCL2 related protein A1 (BCL2A1) [291], and leukemia-associated genes, again BCL2A1 [192], associated with negative values were projected on the vertical axis.

In summary, PCA showed very similar results to cluster analysis for both gene and protein expression data. Interestingly, grouping of cell lines according to their tissue of origin was more homogenous based on gene than by protein expression data. PCA showed inconsistent clustering based on protein expression. Only in the case of PCA_{DE50}, the leukemia cancer cells were clearly separated from the other cancer types. Thus, PCA showed considerably higher degree of divergence among gene and protein expression data because the separation of the individual cancers, as well as the identification of the mostly influencing genes between the transcriptome and the proteome level clearly differed.

4.4.2 CA from all available genes and proteins and from the union of top 50 DE genes and proteins

In this study CA was applied for the first time on gene and protein expression data of the NCI-60 cancer cell lines. Therefore, the findings of the CA can only be compared to the results of the cluster analysis and PCA.

Both for CA from all available genes and proteins (CA_{ALL}; Figure 14) as well, as from the union of top 50 DE genes and proteins (CA_{DE50}; Figure 15), it could be seen that the first two dimensions in gene expression level explained a larger proportion of variance (CA_{ALL}: Dimension 1 (D1): 12% and Dimension 2 (D2): 10.2% and CA_{DE50}: D1: 23.1% and D2: 17.5%) compared to these in protein expression patterns (CA_{ALL}: D1: 4.7% and D2: 3.8% and CA_{DE50}: D1: 13.2% and D2: 8.4%). Furthermore, it could be observed that cancer cell lines were more homogenously clustered according to their tissue origin on the proteome level in contrast to the transcriptome level (Figures 14 – 15). It should be noted that the CA_{DE50} (Figure 17) led again to a better clustering than in the CA_{ALL} (Figure 14).

On the transcriptional level, the melanoma cancer cell lines clustered well (eight out of nine). The outlier was again ME: LOXIMVI, which lacks melanin production [275,276] (Figures 14 – 15). Cancer cell lines, belonging to leukemia, colon and CNS, were grouped into individual clusters (Figures 14 – 15). These findings had also been observed in other studies

which showed that these cancer cell lines were grouped with respect to their tissue origin [269–274]. Further, it could be seen that the cancer cell lines of breast, lung, ovarian and prostate tissue were scattered over various inhomogeneous clusters (Figures 14 – 15). Clustering of luminal and basal breast cancer cell lines was consistent with the results in the cluster analysis and PCA. The two ALL cell lines which are close to each other on the transcriptome level showed the same behavior in PCA.

On the proteome level all cancer cell lines showed homogenous clustering results (Figures 14 – 15) but it is worth mentioning that clustering was more pronounced for the CA_{DE50} (Figure 15) compared to the CA_{ALL} based on protein expression (Figure 14). The findings of the CA concerning all breast cancer cell lines and the two leukemia cancer cell lines were consistent with the results in the CA of the gene expression data.

Based on gene expression of CA_{ALL} the mesenchymal markers, such as SPARC, and collagen markers, including Collagen Type I Alpha 2 (COL1A2), collagen type VIII alpha 1 chain (COL8A1), Collagen Type V Alpha 2 (COL5A1) and collagen type I alpha 1 chain (COL1A1), were separated on D1 from colon and breast cancer cell lines associated with MYB, CORO1A and RAB25, member RAS oncogene family (RAB25) [293], as well as epithelial markers, such as S100P and RAB25 [294], and from leukemia-related genes, including protein tyrosine phosphatase, receptor type C (PTPRC) [295] and HCLS1. Melanoma-related genes, such as, MLANA, DCT and TYR, based on gene expression of CA_{ALL} were projected on the positive side of D2 and therefore these cell lines were separated from the other carcinomas on the vertical axis.

CA_{DE50} selected the most influential genes associated with positive values of D1, including S100 calcium binding protein A14 (S100A14), EPCAM and CDH1, which can be regarded as epithelial markers. The mesenchymal markers, such as SPARC, and collagen markers, including COL1A2, associated with negative values were projected on D1. The epithelial and mesenchymal features were separated along the horizontal axis. Leukemia and melanoma cancer cells were distinct from the epithelial and mesenchymal cancers and these cell lines were projected towards one end of D2. Therefore, melanoma genes, such as MLANA and BCL2A1, as well as leukemia-related genes, including HCLS1 and BCL2A1, accumulated with positive values on D2.

Based on protein expression CA_{50DE} separated the three breast cancer cells BR:HS578T, BR:BT_549 and BR:MDA_MB_231, as well as the CNS, lung, melanoma, ovarian, prostate and

renal carcinomas clearly with positive values on D1. Based on protein expression CA_{50DE} showed that the leukemia, colon and the two breast cancer cells BR:MCF7 and BR:T47D were separated from the other carcinomas on D2. Leukemia-related genes, such as NCK associated protein 1 like (NCKAP1L) [296], also known as HEM1, HCLS1, PTPRC and dedicator of cytokinesis 2 (DOCK2) [297], and colon associated genes, including cyclin dependent kinase inhibitor 2A (CDKN2A) [298,299], were thus projected with positive values on D2. CA_{ALL} based on protein expression resulted in slightly more inconsistent clustering compared to CA_{50DE}. The most influential genes at each end of D1 and D2 did also not provide any real information regarding the clustering of the individual cancers compared to CA_{50DE}.

In summary, it could be noted that the findings of the results of CA were in strong agreement with the results of PCA, as well as cluster analysis, according to gene expression. It is noteably that CA showed homogenous clustering based on protein expression compared to cluster analysis, PCA and NSCA.

4.4.3 NSCA from all available genes and proteins and from the union of top 50 DE genes and proteins

In this study NSCA was applied for the first time on gene and protein expression data of the NCI-60 cancer cell lines. Therefore, the results of the NSCA can only be compared with the previous findings of the cluster analysis, PCA and CA.

For NSCA from all available genes and proteins ($NSCA_{ALL}$; Figure 16), as well as from the union of top 50 DE genes and proteins ($NSCA_{DE50}$; Figure 17), it could be shown that the first two dimensions in gene expression data represented a higher proportion of variance ($NSCA_{ALL}$: D1: 12% and D2: 10.2% and $NSCA_{DE50}$: D1: 23.5% and D2: 17%) compared to these in protein expression data ($NSCA_{ALL}$: D1: 4.6% and D2: 3.9% $NSCA_{DE50}$: D1: 11% and D2: 7.7%). Furthermore, it could be noted that the cancer cell lines in NSCA were more homogenously clustered according to their tissue origin on the transcriptome level compared to the proteome level (Figures 16 – 17). The $NSCA_{DE50}$ (Figure 17) led to a clearer separation of the cancer cell lines in contrast to the $NSCA_{ALL}$ (Figure 16).

On the transcriptome level, the melanoma and leukemia cancer cell lines clustered well and these cancer cell lines clearly differed from the other tissue categories (Figures 16 – 17). Once again, the melanoma cancer cell line, ME:LOXIMVI, was established as an outlier. Cancer cell lines originating from colon and CNS were projected close to each other which determined their tissue of origin (Figures 16 – 17). Other studies led to similar results which showed that these cancer

cell lines were localized according to their tissue category [269–274]. All further cancer cell lines belonging to breast, lung, ovarian, prostate and renal tissue were distributed over several clusters (Figures 16 – 17). Clustering of luminal and basal breast cancer cell lines was consistent with the results in the cluster analysis, PCA and CA. Furthermore, the ER– breast cancer cell lines were again localized closely to CNS malignancies. ER+ breast cancer lines were projected close to colon cancer lines. Similar findings were also confirmed in previous studies [141]. The two ALL cell lines, which were projected close to each other, led to the same results as in PCA and CA.

On D1 based on gene expression of $NSCA_{ALL}$ the mesenchymal markers, such as SPARC, and collagen markers, including COL1A2, were separated from colon and breast cancer cell lines associated with MYB, RAB25 and CORO1A, as well as epithelial markers, such as RAB25 and S100P, and from leukemia-related genes, including HCLS1 and PTPRC. Melanoma-related genes, such as MLANA, DCT and TYR, based on gene expression of $NSCA_{ALL}$ were projected on the positive side of D2 and therefore these cell lines were separated on the vertical axis from the other carcinomas.

$NSCA_{DE50}$ selected the most influential genes associated with positive values of D1, including S100A14, EPCAM and CDH1, which can be regarded as epithelial markers. The mesenchymal markers, such as SPARC, and collagen markers, including COL1A2, associated with negative values were projected on D1. The epithelial and mesenchymal features were separated along the horizontal axis. Leukemia and melanoma cancer cells were distinct from the epithelial and mesenchymal cancers and these cell lines were projected towards one end of D2. Therefore, melanoma genes, such as MLANA, BCL2A1 and SRY-box 10 (SOX10) [300], as well as leukemia-related genes, including BCL2A1 and HCLS1, accumulated with positive values on D2.

In summary, it can be seen that the results of NSCA were in strong consistency with the results of CA, PCA, as well as cluster analysis, based on gene expression. The NSCA showed inconsistent clustering based on protein expression. Only in the case of $NSCA_{ALL}$, the leukemia and melanoma cancer cells were separated from the other cancer types and $NSCA_{DE50}$ showed a clear separation of the leukemia cancer cells. Thus, NSCA showed considerably higher degree of divergence among gene and protein expression data because the separation of the individual cancers, as well as the identification of the mostly influencing genes between the transcriptome and the proteome level clearly differed.

4.5 Integrative analysis

Integrative analysis methods are able to integrate multiple data sets in a meaningful way [88]. Analysis of multiple omics data are performed simultaneously and should therefore allow a deeper insight and comprehensive understanding of a biological system. Several different integrative analysis methods have been developed, such as RGCCA [118], SGCCA [119], penalized canonical correlation analysis [120] and consensus PCA [121], which facilitate the integration of more than two data sets and enable the simultaneous analysis of multiple omics data sets. In this study, MCIA was used as integrative analysis method on gene and protein expression data of the NCI-60 cancer cell lines [86,88].

4.5.1 MCIA from all available genes and proteins and from the union of top 50 DE genes and proteins

MCIA from all available genes and proteins ($MCIA_{ALL}$), as well as from the union of top 50 DE genes and proteins ($MCIA_{DE50}$), revealed similar trends between transcriptome and proteome patterns, indicating that cancer cell lines derived from the same tissue origin were projected into clusters or close to each other (Figures 18 – 19). D1 of $MCIA_{ALL}$ explained 16.1% variance, whereas D2 accounted for 13.6% variance (Figure 18). D1 of $MCIA_{DE50}$ explained 33.2% variance and D2 accounted for 24% variance (Figure 19). $MCIA_{ALL}$ and $MCIA_{DE50}$ clearly explained a higher percentage of variance compared to PCA, CA and NSCA and therefore, the internal structure of the data was best explained and revealed by MCIA. This is also the reason why MCIA resulted in a clearer separation of the cancer cell lines compared to PCA, CA and NSCA.

The leukemia, melanoma, CNS, renal and ovarian cancer cell lines were largely grouped by their tissue origin, similar to a previous finding [44,86,112], whereas cell lines, originating from breast, lung and prostate carcinomas with higher intrinsic molecular heterogeneity, were not clustered homogenously, consistent with results of a prior study (Figures 18 – 19) [44,86,112]. The ER+ breast cancer lines (BR:MCF7 and BR:T47D) showed epithelial phenotypes and were closely clustered with colon cancer lines. ER– breast cancer cell lines (BR:HS578T and BR:BT_549) were grouped geometrically close with stromal/mesenchymal cluster of glioblastoma and renal cancer cell lines. Thus, the ER– breast cancer cell lines were likely to have more invasive mesenchymal features compared to ER+ breast cancer lines. These findings agree with previous immunohistochemical studies based on breast cancer cells [301]. One

melanoma cancer cell line, ME:LOXIMVI, was projected outside the melanoma cluster. This cancer cell line lacks melanin and produces no characteristic melanocyte genes [277,278].

However, there were also some cancer cell lines which showed a greater divergence between the gene and the protein expression profiles, which resulted in a lower correlation between their transcriptome and proteome level. The projection of the melanoma cancer cell line, ME:SK_MEL_2, according to protein expression was further away from the other melanoma cancer cell lines and had a smaller distance to the plot origin compared to the same cancer cell type in the gene expression data set (Figures 18 – 19). This result was also confirmed in [86] and could mean that less melanin-related genes are expressed on protein levels. Furthermore, the deviation of the ovarian cancer cell line, OV:IGROV1, could result from the expression of less epithelial markers which were displayed on the positive direction of the second dimension (Figure 19), similar to a previous finding [86]. Such a discrepancy can also point to biological variance, a batch effect or a technical artifact. Thus, a significant advantage of the MCIA is the ability to determine the presence or absence of co-structure among data sets. Therefore, the strongest features of each data set can be identified and selected for further analysis.

The integrative analysis of transcriptome and proteome profiles (MCIA) based on $MCIA_{ALL}$ revealed some colon and breast cancer cells associated genes, like MYB and CORO1A, leukemia-related genes, such as PTPRC, also known CD45 antigen (CD45) [302] and HCLS1, and epithelial markers, including S100P, all associated with positive values on the D1. The PTPRC and HCLS1 genes are involved in different processes in immune cells [295,303]. Collagen markers, such as COL1A2, and mesenchymal markers, including SPARC, were projected with negative values on the D1. Collagen markers are regarded as the main components of extracellular matrix of connective tissue [304] and are required for metastasis of cancer cell lines [304]. Furthermore, they influence the invasive, metastatic phenotype in cancer cells [305,306] and therefore also promote the epithelial-mesenchymal transition (EMT) which has a significant function in cancer biology and is also involved the malignancy and metastasis of epithelial cancer cells including breast and colon carcinomas [301,307,308]. SPARC can also be associated with collagens, extracellular matrix and EMT [309]. SPARC is used to calcify the collagen in the bone, for the extracellular matrix synthesis and for the support of changes to cell shape [310]. Furthermore, it plays an important role in tumor suppression, as well as it influence metastasis due to the cell shape changes which can support tumor cell invasion [310]. Thus, the colon, breast and leukemia cancer cells, as well as the epithelial markers, were separated from the mesenchymal features (glioblastoma and renal cancer cell lines) along the D1. The positive projections of D2 based on $MCIA_{ALL}$ revealed some melanocyte-specific genes, including

S100B, TYR, DCT and MLANA. These proteins were not among the most influential genes in the protein expression data and this is the reason why the melanoma cancer cell line, ME:LOXIMVI, which lacks melanin, was projected closer to the plot origin in the protein expression space. Thus, the melanoma cancer cells were clearly separated from the other cancer tissues and thus melanoma cell lines were projected on the positive side of the D2.

$MCIA_{DE50}$ showed a number of collagen genes, such as COL1A2, COL5A1 and Collagen Type VI Alpha 2 (COL6A2), and mesenchymal markers, including SPARC, associated with negative values of the D1. It revealed some epithelial markers, such as SLC27A2, CDH1, EPCAM, KRT19 and SPINT2 associated with positive values of the D1 which are similar to the findings in [86,112]. As a result, D1 was associated with the mechanism that considers the variant potential of the metastatic process of cancer cell lines. Furthermore, this horizontal axis clustered cancer cell lines with epithelial (leukemia and colon cancer cells) or mesenchymal (glioblastoma and renal cancer cells) features which also confirmed that EMT plays an important role in various cancer types. Mesenchymal cells have migratory and invasive characteristics which have a significant function in malignant metastasizing cancer and therefore, EMT targets in investigating the biology of different cancer classes [307]. $MCIA_{DE50}$ showed that the leukemia and melanoma cancer cell lines were projected on the negative side of the D2. HUGO gene symbols of $MCIA_{DE50}$ based on transcriptome data with high weights in the D2 revealed several melanocyte-specific genes, including MLANA, BCL2A1 and SOX10, whereas no leukemia-related markers were found in the most influential genes. Similar results regarding the melanoma cancer cell lines are obtained in [311], which stated that the weights of melanoma cancer cell lines are largely reflected by the transcriptome data. This suggests why the melanoma cancer cell line, ME:LOXIMVI, which lacks melanin, was projected closer to the plot origin in the protein expression space. $MCIA_{DE50}$ based on protein expression highlighted several melanoma genes, including MLANA, RAB38, member RAS oncogene family (RAB38) [312] and SOX10, as well as some leukemia-related genes, such as HCLS1 and PTPRC. Leukemia and melanoma cancer cells were distinct from the epithelial and mesenchymal cancers and these cell lines were projected towards one end of the D2.

4.6 Comparison between sequential and integrative analysis

An important feature of MCIA is the projection of the most influential variables (genes). Variables and observations (samples) which are represented along the same direction from the plot origin on the PCs show a high correlation (increased/upregulated in available observations).

Variables on the opposite direction from the plot origin to those observations show a weak association (decreased/downregulated in available observations). Therefore, the most influential variables on any dimension can be selected and these variables should simplify the biological interpretation of PCs. Thus, the projection direction (positive, negative) has no relevance to the results, but rather which variables are projected in the same direction near to each other. These variables, which are clustered close together, thus show a strong correlation and association to one another.

The results regarding the ER+ (BR:MCF7 and BR:T47D) and ER– (BR:HS578T and BR:BT_549) breast cancer cell lines, as well as the findings of the melanoma cancer cell line ME:LOXIMVI, obtained in the MCIA were consistent with the findings of the above-mentioned sequential analysis methods (PCA, CA, NSCA) and were also confirmed by the results of a previous study [86].

$MCIA_{ALL}$ separated the mesenchymal markers, such as SPARC, and the collagen markers, like COL1A2, from the colon and breast cancer cells associated with genes, like MYB and CORO1A, leukemia-related genes, including PTPRC and HCLS1, and epithelial markers, such as S100P, on D1. Melanoma markers, such as MLANA, DCT, TYR and S100B, were separated from the other carcinomas along D2. The three sequential analytical methods (PCA_{ALL}, CA_{ALL}, $NSCA_{ALL}$) showed that the separation of the most important genes based on transcriptome data on PC1/D1 and PC2/D2 was similar to $MCIA_{ALL}$ (Table 21). In all four methods (PCA_{ALL}; CA_{ALL}, $NSCA_{ALL}$, $MCIA_{ALL}$) the mesenchymal markers are separated from the colon, breast, leukemia cancer cells, as well as the epithelial marker along PC1/D1 and the melanoma cancer cells were separated from the other carcinomas on PC2/D2 (Table 21).

$MCIA_{DE50}$ separated the mesenchymal features, such as SPARC, and the collagen markers, including COL1A2, COL5A1 and COL6A2, from the epithelial markers, such as CDH1, EPCAM, KRT19, SPINT2 and SLC27A2, on D1. The three sequential analytical methods (PCA_{50DE}, CA_{50DE}, $NSCA_{50DE}$) showed that the separation of the most important genes based on transcriptome data on PC1/D1 and PC2/D2 was similar to $MCIA_{DE50}$ (Table 22). In all four methods (PCA_{50DE}; CA_{50DE}, $NSCA_{50DE}$, $MCIA_{50DE}$) the mesenchymal markers are separated from the epithelial markers along PC1/D1 and the melanoma and leukemia cancer cells were separated from the other carcinomas on PC2/D2 (Table 22).

PCA and NSCA showed inconsistent clustering based on protein expression. Only in the case of PCA_{DE50}, the leukemia cancer cells were clearly separated from the other cancer types. In the

case of $NSCA_{ALL}$, the leukemia and melanoma cancer cells were separated from the other cancer types and $NSCA_{DE50}$ showed a clear separation of the leukemia cancer cells. Thus, PCA and NSCA showed considerably higher degree of divergence among gene and protein expression data because the separation of the individual cancers, as well as the identification of the mostly influencing genes between the transcriptome and the proteome level clearly differed. It is noteably that CA showed homogenous clustering based on protein expression compared to PCA and NSCA. This is because there was a significant dependency between the rows and columns according to the proteome profile. A method to examine the association between rows and columns is the chi-square statistics [313]. It follows that the chi-square statistics being highly significant based on the proteome data (CA_{ALL}: chi-square = 1262542 and p-value = 0; CA_{50DE}: chi-square = 81560.96 and p-value = 0). Thus, the chi-square statistics based on protein expression data specified a strong association between row and column variables. This suggests that the CA provides a more homogeneous separation of the carcinomas based on protein expression and thus, also provides a different scaling with respect to the x-axis as opposed to the other sequential analysis methods. $MCIA_{ALL}$ based on protein expression provided also homogenous clusters of the different cancer types.

Protein and gene expression profiles can be regarded as powerful molecular descriptors of different cancer tissues and integrative analysis of both provides a deeper insight into multiple layers of biological systems compared to any single analysis tool. As expected, the most homogenous clustering and separation of the different carcinomas according to the most influential genes was achieved with the MCIA, and thus, the different types of cancer were clearly separated on the two PCs by the integrative analysis method. The most influential genes based on MCIA at each end of PC1 and PC2 provided useful information regarding the clustering of the individual cancers. Furthermore, MCIA achieves a homogenous separation of mesenchymal (glioblastoma and renal cancer cells) and epithelial (leukemia and colon cancer cells) markers which promote EMT that has a significant function in cancer biology and is also involved the malignancy and metastasis of epithelial cancer cells [301,307,308]. Mesenchymal cells have migratory and invasive characteristics which have a significant function in malignant metastasizing cancer and therefore. Thus, MCIA also yields important information regarding invasive, metastatic phenotype in cancer cells [305,306].

5 Literature

1. International Human Genome Sequencing Consortium: **Finishing the euchromatic sequence of the human genome**. *Nature* 2004, **431**:931–945.
2. McDermott JE, Wang J, Mitchell H, Webb-Robertson B-J, et al.: **Challenges in Biomarker Discovery: Combining Expert Insights with Statistical Analysis of Complex Omics Data**. *Expert Opinion on Medical Diagnostics* 2013, **7**:37–51.
3. Zhang W, Li F, Nie L: **Integrating multiple "omics" analysis for microbial biology: application and methodologies.** *Microbiology* 2010, **156**:287–301.
4. World Health Organization. **WHO definitions of genetics and genomics**. http://www.who.int/genomics/geneticsVSgenomics/en. 2016-01-25.
5. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Research* 2005, **33**:501–504.
6. Hubbard T, Barker D, Birney E, Cameron G, et al.: **The Ensembl genome database project**. *Nucleic Acids Research* 2002, **30**:38–41.
7. Campo E: **Whole genome profiling and other high throughput technologies in lymphoid neoplasms-current contributions and future hopes**. *Modern Pathology* 2013, **26**:97–110.
8. Bejjani BA, Shaffer LG: **Application of Array-Based Comparative Genomic Hybridization to Clinical Diagnostics**. *The Journal of Molecular Diagnostics* 2006, **8**:528–533.
9. Zhang F, Gu W, Hurles ME, Lupski JR: **Copy Number Variation in Human Health, Disease, and Evolution**. *Annual Review of Genomics and Human Genetics* 2009, **10**:451–481.
10. Pinkel D, Albertson DG: **Array comparative genomic hybridization and its applications in cancer**. *Nature Genetics* 2005, **37**:Suppl:S11-7.
11. LaFramboise T: **Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances**. *Nucleic Acids Research* 2009, **37**:4181–4193.
12. Zhao X, Li C, Paez JG, Chin K, et al.: **An Integrated View of Copy Number and Allelic Alterations in the Cancer Genome Using Single Nucleotide Polymorphism Arrays**. *Cancer Research* 2004, **64**:3060–3071.
13. Mullighan CG, Goorha S, Radtke I, Miller CB, et al.: **Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia**. *Nature* 2007, **446**:758–764.
14. Pareek CS, Smoczynski R, Tretyn A: **Sequencing technologies and genome sequencing**. *Journal of Applied Genetics* 2011, **52**:413–435.
15. Lodish HF, Berk A, Kaiser C, Krieger M, et al.: *Molecular Cell Biology*. 7 edn. New York: W. H.. Freeman and Company; 2012.
16. Bird A: **DNA methylation patterns and epigenetic memory**. *Genes & Development* 2002, **16**:6–21.
17. Kouzarides T: **Chromatin Modifications and Their Function**. *Cell* 2016, **128**:693–705.
18. Strahl BD, Allis CD: **The language of covalent histone modifications**. *Nature* 2000, **403**:41–45.
19. Meissner A: **Epigenetic modifications in pluripotent and differentiated cells**. *Nature Biotechnology* 2010, **28**:1079–1088.
20. Zhu J, Adli M, Zou JY, Verstappen G, et al.: **Genome-wide chromatin state transitions associated with developmental and environmental cues**. *Cell* 2013, **152**:642–654.
21. Holliday R, Jeggo PA: **Mechanisms for changing gene expression and their possible relationship to carcinogenesis**. *Cancer Surveys* 1985, **4**:557–581.
22. Alegría-Torres JA, Baccarelli A, Bollati V: **Epigenetics and lifestyle**. *Epigenomics* 2011, **3**:267–277.
23. Feinberg AP, Tycko B: **The history of cancer epigenetics**. *Nature Reviews Cancer* 2004, **4**:143–153.
24. Esteller M: **Aberrant DNA methylation as a cancer-inducing mechanism**. *Annual Review of Pharmacology and Toxicology* 2005, **45**:629–656.
25. Frommer M, McDonald LE, Millar DS, Collis CM, et al.: **A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands.** *Proceedings of the National Academy of Sciences of the United States of America* 1992, **89**:1827–1831.
26. Li Y, Tollefsbol TO: **DNA methylation detection: Bisulfite genomic sequencing analysis**. *Methods in Molecular Biology* 2011, **791**:11–21.
27. Rand K, Qu W, Ho T, Clark SJ, et al.: **Conversion-specific detection of DNA methylation using real-time polymerase chain reaction (ConLight-MSP) to avoid false positives**. *Methods* 2002, **27**:114–120.
28. Xiong Z, Laird PW: **COBRA: a sensitive and quantitative DNA methylation assay.** *Nucleic Acids Research* 1997, **25**:2532–2534.
29. Gonzalgo ML, Jones PA: **Rapid quantitation of methylation differences at specific sites using methylation-sensitive single nucleotide primer extension (Ms-SNuPE).** *Nucleic Acids Research* 1997, **25**:2529–2531.
30. Carey MF, Peterson CL, Smale ST: **Chromatin Immunoprecipitation (ChIP)**. *Cold Spring Harbor Protocols* 2009, **2009**:pdb.prot5279.
31. Kuo M-H, Allis CD: **In Vivo Cross-Linking and Immunoprecipitation for Studying Dynamic Protein:DNA Associations in a Chromatin Environment**. *Methods* 1999, **19**:425–433.
32. Buck MJ, Lieb JD: **ChIP-chip: considerations for the design, analysis, and application of genome-wide**

chromatin immunoprecipitation experiments. *Genomics* 2004, **83**:349–360.
33. Landt SG, Marinov GK, Kundaje A, Kheradpour P, et al.: **ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia**. *Genome Research* 2012, **22**:1813–1831.
34. Horgan RP, Kenny LC: **"Omic" technologies: genomics, transcriptomics, proteomics and metabolomics**. *The Obstetrician & Gynaecologist* 2011, **13**:189–195.
35. Blow N: **Transcriptomics: The digital generation**. *Nature* 2009, **458**:239–242.
36. Mattick JS, Makunin I: **Non-coding RNA**. *Human Molecular Genetics* 2006, **15**:17–29.
37. Lodish H, Berk A, Zipursky SL: **The Three Roles of RNA in Protein Synthesis**. In *Molecular Cell Biology*. Edited by Ahr K. New York: W. H. Freeman; 2000:119–124.
38. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, et al.: **miRBase: microRNA sequences, targets and gene nomenclature**. *Nucleic Acids Research* 2006, **34**:D140–D144.
39. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nature Reviews Genetics* 2009, **10**:57–63.
40. Dong Z, Chen Y: **Transcriptomics: Advances and approaches**. *Science China Life Sciences* 2013, **56**:960–967.
41. Rhodes DR, Chinnaiyan AM: **Integrative analysis of the cancer transcriptome**. *Nature Genetics* 2005, **37**:31–37.
42. Butte AJ, Tamayo P, Slonim D, Golub TR, et al.: **Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks**. *Proceedings of the National Academy of Sciences* 2000, **97**:12182–12186.
43. Staunton JE, Slonim DK, Coller HA, Tamayo P, et al.: **Chemosensitivity prediction by transcriptional profiling**. *Proceedings of the National Academy of Sciences* 2001, **98**:10787–10792.
44. Ross DT, Scherf U, Eisen MB, Perou CM, et al.: **Systematic variation in gene expression patterns in human cancer cell lines**. *Nature Genetics* 2000, **24**:227–235.
45. Kagnoff MF, Eckmann L: **Analysis of host responses to microbial infection using gene expression profiling**. *Current Opinion in Microbiology* 2001, **4**:246–250.
46. Woo Y, Affourtit J, Daigle S, Viale A, et al.: **A Comparison of cDNA, Oligonucleotide, and Affymetrix GeneChip Gene Expression Microarray Platforms**. *Journal of Biomolecular Techniques* 2004, **15**:276–284.
47. Liang M, Cowley AW, Greene AS: **High throughput gene expression profiling: a molecular approach to integrative physiology**. *The Journal of Physiology* 2004, **554**:22–30.
48. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proceedings of the National Academy of Sciences of the United States of America* 1977, **74**:5463–5467.
49. Boguski MS, Tolstoshev CM, Bassett DE: **Gene Discovery in dbEST**. *Science* 1994, **265**:1993–1994.
50. Zhang J, Chiodini R, Badr A, Zhang G: **The impact of next-generation sequencing on genomics**. *Journal of Genetics and Genomics* 2011, **38**:95–109.
51. Harbers M, Carninci P: **Tag-based approaches for transcriptome research and genome annotation**. *Nature Methods* 2005, **2**:495–502.
52. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial Analysis of Gene Expression**. *Science* 1995, **270**:484–487.
53. Kodzius R, Kojima M, Nishiyori H, Nakamura M, et al.: **CAGE: cap analysis of gene expression**. *Nature Methods* 2006, **3**:211–222.
54. Reinartz J, Bruyns E, Lin J-Z, Burcham T, et al.: **Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms**. *Briefings in Functional Genomics & Proteomics* 2002, **1**:95–104.
55. McGettigan PA: **Transcriptomics in the RNA-seq era**. *Current Opinion in Chemical Biology* 2013, **17**:4–11.
56. de Hoog CL, Mann M: **Proteomics**. *Annual Review of Genomics and Human Genetics* 2004, **5**:267–293.
57. Patterson SD, Aebersold RH: **Proteomics: the first decade and beyond**. *Nature Genetics* 2003, **33**:311–323.
58. Graves PR, Haystead TAJ: **Molecular Biologist's Guide to Proteomics**. *Microbiology and Molecular Biology Reviews* 2002, **66**:39–63.
59. Mann M, Jensen ON: **Proteomic analysis of post-translational modifications**. *Nature Biotechnology* 2003, **21**:255–261.
60. Gygi SP, Corthals GL, Zhang Y, Rochon Y, et al.: **Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology**. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**:9390–9395.
61. Rabilloud T, Lelong C: **Two-dimensional gel electrophoresis in proteomics: A tutorial**. *Journal of Proteomics* 2011, **74**:1829–1841.
62. Mann M, Hendrickson RC, Pandey A: **Analysis of Proteins and Proteomes by Mass Spectrometry**. *Annual Review of Biochemistry* 2001, **70**:437–473.
63. Lin D, Tabb DL, Yates III JR: **Large-scale protein identification using mass spectrometry**. *Biochimica et Biophysica Acta* 2003, **1646**:1–10.
64. Zhang G, Ueberheide BM, Waldemarson S, Myung S, et al.: **Protein Quantitation Using Mass Spectrometry**. *Methods in Molecular Biology* 2010, **673**:211–222.
65. Smith ICP, Blandford DE: **Nuclear magnetic resonance spectroscopy**. *Analytical Chemistry* 1995, **67**:509–

518.
66. Schmidt D, Wilson MD, Spyrou C, Brown GD, et al.: **ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions**. *Methods* 2009, **48**:240–248.
67. Frangioni J V: **In vivo near-infrared fluorescence imaging**. *Current Opinion in Chemical Biology* 2003, **7**:626–634.
68. Kiemer L, Cesareni G: **Comparative interactomics: comparing apples and pears?** *Trends in Biotechnology* 2016, **25**:448–454.
69. Serebriiskii I: **Yeast Two-Hybrid System for Studying Protein-Protein Interactions--Stage 2: Transforming and Characterizing the Library**. *Cold Spring Harbor Protocols* 2010, **2010**:pdb.prot5430.
70. Brückner A, Polge C, Lentze N, Auerbach D, et al.: **Yeast Two-Hybrid, a Powerful Tool for Systems Biology**. *International Journal of Molecular Sciences* 2009, **10**:2763–2788.
71. Hall DA, Ptacek J, Snyder M: **Protein Microarray Technology**. *Mechanisms of Ageing and Development* 2007, **128**:161–167.
72. Demain AL: **Microbial production of primary metabolites**. *Naturwissenschaften* 1980, **67**:582–587.
73. Büschl C: **Novel software for stable isotopic labelling assisted and LC-HRMS based untargeted metabolomics research and tracer-fate studies**. Dissertation. Graz University of Technology. Institute for Knowledge Discovery. 2015.
74. Roessner U, Bowne J: **What is metabolomics all about?** *Biotechniques* 2009, **46**:363–365.
75. Cates VE, Meloan CE: **Compound identification by gas chromatography**. *Journal of Chromatography A* 1963, **12**:15–21.
76. Zhou B, Xiao JF, Tuli L, Ressom HW: **LC-MS-based metabolomics**. *Molecular BioSystems* 2012, **8**:470–481.
77. Gika HG, Theodoridis GA, Wingate JE, Wilson ID: **Within-Day Reproducibility of an HPLC−MS-Based Method for Metabonomic Analysis: Application to Human Urine**. *Journal of Proteome Research* 2007, **6**:3291–3303.
78. Nováková L, Matysová L, Solich P: **Advantages of application of UPLC in pharmaceutical analysis**. *Talanta* 2006, **68**:908–918.
79. Soga T, Ohashi Y, Ueno Y, Naraoka H, et al.: **Quantitative Metabolome Analysis Using Capillary Electrophoresis Mass Spectrometry**. *Journal of Proteome Research* 2003, **2**:488–494.
80. Dettmer K, Aronov PA, Hammock BD: **Mass Spectrometry-Based Metabolomics**. *Mass Spectrometry Reviews* 2007, **26**:51–78.
81. Jie MS, Mustafa J: **High-resolution nuclear magnetic resonance spectroscopy--applications to fatty acids and triacylglycerols**. *Lipids* 1997, **32**:1019–1034.
82. Romero R, Espinoza J, Gotsch F, Kusanovic JP, et al.: **The use of high-dimensional biology (genomics, transcriptomics, proteomics, and metabolomics) to understand the preterm parturition syndrome**. *BJOG: An International Journal of Obstetrics & Gynaecology* 2006, **113**:118–135.
83. Lê Cao K-A, González I, Déjean S: **integrOmics: an R package to unravel relationships between two omics datasets**. *Bioinformatics* 2009, **25**:2855–2856.
84. Meng C, Culhane A, Gholami AM. **Package 'omicade4'**. https://www.bioconductor.org/packages/release/bioc/manuals/omicade4/man/omicade4.pdf. Bioconductor project. 2016-02-15.
85. Meng C, Gholami AM. **Multiple Co-inertia Analysis of Multiple OMICS Data using omicade4**. https://www.bioconductor.org/packages/3.3/bioc/vignettes/omicade4/inst/doc/omicade4.pdf. Bioconductor project. 2016-02-15.
86. Meng C, Kuster B, Culhane AC, Gholami AM: **A multivariate approach to the integration of multi-omics datasets**. *BMC Bioinformatics* 2014, **15**:1–13.
87. Mardis ER: **The $1,000 genome, the $100,000 analysis?** *Genome Medicine* 2010, **2**:84.
88. Meng C, Zeleznik OA, Thallinger GG, Kuster B, et al.: **Dimension reduction techniques for the integrative analysis of multi-omics data**. *Briefings in Bioinformatics* 2016:in press.
89. Fagan A, Culhane AC, Higgins DG: **A multivariate analysis approach to the integration of proteomic and gene expression data**. *Proteomics* 2007, **7**:2162–2171.
90. Alter O, Brown PO, Botstein D: **Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**:3351–3356.
91. Tomescu OA, Mattanovich D, Thallinger GG: **Integrative omics analysis. A study based on *Plasmodium falciparum* mRNA and protein data**. *BMC Systems Biology* 2014, **8**:S4.
92. Cox B, Kislinger T, Emili A: **Integrating gene and protein expression data: pattern analysis and profile mining**. *Methods* 2005, **35**:303–314.
93. Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, Grainger M, Yan SF, Williamson KC, Holder A, Carucci DJ, Yates JR, Winzeler E: **Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle**. *Genome Research* 2004, **14**:2308–2318.
94. Wang KS, Liu X: **Integrative Analysis of Genome-wide Expression and Methylation Data**. *Journal of Biometrics & Biostatistics* 2013, **4**:4–6.
95. Chen Z, Zhang W: **Integrative Analysis Using Module-Guided Random Forests Reveals Correlated**

Genetic Factors Related to Mouse Weight. *PLoS Computational Biology* 2013, **9**:e1002956.
96. Mo Q, Wang S, Seshan VE, Olshen AB, et: **Pattern discovery and cancer gene identification in integrated cancer genomic data**. *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**:4245–4250.
97. Kockmann T, Gerstung M, Schlumpf T, Xhinzhou Z, et al.: **The BET protein FSH functionally interacts with ASH1 to orchestrate global gene activity in Drosophila**. *Genome Biology* 2013, **14**:R18.
98. Gerstein MB, Kundaje A, Hariharan M, Landt SG, et: **Architecture of the human regulatory network derived from ENCODE data**. *Nature* 2012, **489**:91–100.
99. Tebaldi T, Re A, Viero G, Pegoretti I, et al.: **Widespread uncoupling between transcriptome and translatome variations after a stimulus in mammalian cells**. *BMC Genomics* 2012, **13**:220.
100. Jolliffe IT: *Principal Component Analysis*. 2 edn. New York, NY, USA: Springer-Verlag; 2002.
101. Raychaudhuri S, Stuart JM, Altman RB: **Principal components analysis to summarize microarray experiments: application to sporulation time series**. *Proceedings of the Pacific Symposium on Biocomputing* 2000:455–466.
102. Yeung KY, Ruzzo WL: **Principal component analysis for clustering gene expression data**. *Bioinformatics* 2001, **17**:763–774.
103. Yao F, Coquery J, Lê Cao K-A: **Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets**. *BMC Bioinformatics* 2012, **13**:24.
104. Greenacre MJ: *Theory and Applications of Correspondence Analysis*. 1 edn. London: Academic Press; 1984.
105. Fellenberg K, Hauser NC, Brors B, Neutzner A, et al.: **Correspondence analysis applied to microarray data**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**:10781–10786.
106. Beh EJ, Lombardo R: *Correspondence Analysis: Theory, Practice and New Strategies*. 1 edn. United Kingdom: John Wiley & Sons; 2014.
107. Greenacre MJ: *Correspondence Analysis in Practice*. 2 edn. Boca Raton, London, New York: Taylor & Francis, Ltd; 2007.
108. Gimaret-Carpentier C, Chessel D, Pascal J: **Non-symmetric correspondence analysis: an alternative for species occurrences data**. *Plant Ecology* 1998, **138**:97–112.
109. Kroonenberg PM, Lombardo R: **Nonsymmetric Correspondence Analysis: A Tool for Analysing Contingency Tables With a Dependence Structure**. *Multivariate Behavioral Research* 1999, **34**:367–396.
110. Dray S, Chessel D, Thioulouse J: **Co-inertia analysis and the linking of ecological data tables**. *Ecology* 2003, **84**:3078–3089.
111. Dolédec S, Chessel D: **Co-inertia analysis: an alternative method for studying species–environment relationships**. *Freshwater Biology* 1994, **31**:277–294.
112. Culhane AC, Perrière G, Higgins DG: **Cross-platform comparison and visualisation of gene expression data using co-inertia analysis**. *BMC Bioinformatics* 2003, **4**:59.
113. Soneson C, Lilljebjörn H, Fioretos T, Fontes M: **Integrative analysis of gene expression and copy number alterations using canonical correlation analysis**. *BMC Bioinformatics* 2010, **11**:191.
114. González I, Déjean S, Martin PGP, Baccini A: **CCA: An R Package to Extend Canonical Correlation Analysis**. *Journal of Statistical Software* 2008, **23**:1–14.
115. Ter Braak CJF: **The analysis of vegetation-environment relationships by canonical correspondence analysis**. *Vegetatio* 1987, **69**:69–77.
116. Palmer MW: **Putting Things in Even Better Order: The Advantages of Canonical Correspondence Analysis**. *Ecology* 1993, **74**:2215–2230.
117. Ponnapalli SP, Saunders MA, Van Loan CF, Alter O: **A Higher-Order Generalized Singular Value Decomposition for Comparison of Global mRNA Expression from Multiple Organisms**. *PLoS One* 2011, **6**:e28072.
118. Tenenhaus A, Tenenhaus M: **Regularized Generalized Canonical Correlation Analysis**. *Psychometrika* 2011, **76**:257–284.
119. Tenenhaus A, Philippe C, Guillemot V, Le Cao K-A, et: **Variable selection for generalized canonical correlation analysis**. *Biostatistics* 2014, **15**:569–583.
120. Witten DM: **A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis**. *Biostatistics* 2009, **10**:515–534.
121. Hanafi M, Kohler A, Qannari E-M: **Connections between multiple co-inertia analysis and consensus principal component analysis**. *Chemometrics and Intelligent Laboratory Systems* 2011, **106**:37–40.
122. Hua L, Zheng W, Xia H, Zhou P, et: **Integration of multi-microarray datasets to identify chronic obstructive pulmonary disease-related miRNAs**. *Bio-Medical Materials and Engineering* 2015, **26**:1903–1915.
123. Reinhold WC, Sunshine M, Liu H, Varma S, et al.: **CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set**. *Cancer Research* 2012, **72**:3499–3511.
124. Shankavaram UT, Varma S, Kane D, Sunshine M, et al.: **CellMiner: a relational database and query tool for the NCI-60 cancer cell lines**. *BMC Genomics* 2009, **10**:277.
125. Cellminer: **Cellminer homepage**. http://discover.nci.nih.gov/cellminer/home.do. 2016-07-11.

126. Shankavaram UT, Reinhold WC, Nishizuka S, Major S, et al.: **Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study**. *Molecular Cancer Therapeutics* 2007, **6**:820–832.
127. Gene Expression Omnibus. **Platform GPL91**. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL91. 2016-13-07.
128. Gene Expression Omnibus. **Series GSE5949**. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5949. 2016-13-07.
129. Gene Expression Omnibus. **Platform GPL96**. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL96. 2016-13-07.
130. Gene Expression Omnibus. **Series GSE5720**. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5720. 2016-13-07.
131. Gene Expression Omnibus. **Platform GPL570**. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL570. 2016-13-07.
132. Gene Expression Omnibus. **Series GSE32474**. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32474. 2016-13-07.
133. Gene Expression Omnibus. **Platform GPL4133**. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL4133. 2016-13-07.
134. Gene Expression Omnibus. **Series GSE29288**. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE29288. 2016-13-07.
135. Gautier L, Irizarry R, Cope L, Bolstad B: **Description of affy**. https://www.bioconductor.org/packages/devel/bioc/vignettes/affy/inst/doc/affy.pdf. Bioconductor project. 2016-07-11.
136. Irizarry RA: **From CEL Files to Annotated Lists of Interesting Genes**. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Edited by Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S. New York, NY: Springer New York; 2005:431–442.
137. Wu J, Irizarry R: **Package 'gcrma'**. https://www.bioconductor.org/packages/release/bioc/manuals/gcrma/man/gcrma.pdf. Bioconductor project. 2016-07-11.
138. Agilent Technologies. **Agilent GeneSpring**. http://genespring-support.com/files/gs_12_6/GeneSpring-manual.pdf. 2016-07-28.
139. Wu Z, Irizarry R, Gentleman R, Murillo FM, et al.: **A model-based background adjustment for oligonucleotide expression arrays**. *Journal of the American Statistical Association* 2004, **99**:909–917.
140. Feng C, Wang H, Lu N, Chen T, et al.: **Log-transformation and its implications for data analysis**. **Shanghai Archives of Psychiatry** 2014, **26**:105–109.
141. Gholami AM, Hahne H, Wu Z, Auer FJ, et: **Global Proteome Analysis of the NCI-60 Cell Line Panel**. *Cell Reports* 2013, **4**:609–620.
142. Schirle M, Heurtier M-A, Kuster B: **Profiling Core Proteomes of Human Cell Lines by One-dimensional PAGE and Liquid Chromatography-Tandem Mass Spectrometry**. *Molecular & Cellular Proteomics* 2003, **2**:1297–1305.
143. Cox J, Matic I, Hilger M, Nagaraj N, Selbach M, Olsen J V, Mann M: **A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics**. *Nature Protocols* 2009, **4**:698–705.
144. Munich University of Technology: **NCI-60**. http://129.187.44.58:7070/NCI60/main/download. 2016-07-11.
145. Munich University of Technology and SAP SE: **Proteomics DB**. https://www.proteomicsdb.org/. 2016-07-11.
146. Ho D, et al.: **Notepad++ 6.9.2 released**. https://notepad-plus-plus.org/news/notepad-6.9.2-released.html. 2016-08-17.
147. Holman JD, Tabb DL, Mallick P: **Employing ProteoWizard to Convert Raw Mass Spectrometry Data**. *Current Protocols in Bioinformatics* 2014, **46**:13.24.1-13.24.9.
148. Gentleman RC, Carey VJ, Bates DM, Bolstad B, et al.: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biology* 2004, **5**:R80.
149. Carlson M: **Package 'hgu95a.db'**. https://bioconductor.org/packages/release/data/annotation/manuals/hgu95a.db/man/hgu95a.db.pdf. Bioconductor project. 2016-07-11.
150. Carlson M: **Package 'hgu95b.db'**. https://bioconductor.org/packages/release/data/annotation/manuals/hgu95b.db/man/hgu95b.db.pdf. Bioconductor project. 2016-07-11.
151. Carlson M: **Package 'hgu95c.db'**. https://bioconductor.org/packages/release/data/annotation/manuals/hgu95c.db/man/hgu95c.db.pdf. Bioconductor project. 2016-07-11.
152. Carlson M: **Package 'hgu95d.db'**. https://bioconductor.org/packages/release/data/annotation/manuals/hgu95d.db/man/hgu95d.db.pdf. Bioconductor project. 2016-07-11.
153. Carlson M: **Package 'hgu95e.db'**. https://bioconductor.org/packages/release/data/annotation/manuals/hgu95e.db/man/hgu95e.db.pdf.

Bioconductor project. 2016-07-11.
154. Carlson M: **Package 'hgu133a.db'**. https://bioconductor.org/packages/release/data/annotation/manuals/hgu133a.db/man/hgu133a.db.pdf. Bioconductor project. 2016-07-11.
155. Carlson M: **Package 'hgu133b.db'**. https://bioconductor.org/packages/release/data/annotation/manuals/hgu133b.db/man/hgu133b.db.pdf. Bioconductor project. 2016-07-11.
156. Carlson M: **Package 'hgu133plus2.db'**. https://bioconductor.org/packages/release/data/annotation/manuals/hgu133plus2.db/man/hgu133plus2.db.pdf. Bioconductor project. 2016-07-11.
157. Gene Expression Omnibus: **Platform GPL20769**. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL20769. 2016-07-11.
158. Smyth GK: **Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments**. *Statistical Applications in Genetics and Molecular Biology* 2004, **3**:1–25.
159. Ritchie ME, Phipson B, Wu D, Hu Y, et al.: **limma powers differential expression analyses for RNA-sequencing and microarray studies**. *Nucleic Acids Research* 2015, **43**:e47–e47.
160. Smyth G K, Ritchie M, Thorne N, Wettenhall J, et al. **limma: Linear Models for Microarray and RNA-Seq Data User's Guide**. https://www.bioconductor.org/packages/3.3/bioc/vignettes/limma/inst/doc/usersguide.pdf. Bioconductor project. 2016-03-02.
161. Smyth GK: **Limma: linear models for microarray data**. In *In Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Edited by Gentleman R, Carey V, Dudoit S, Irizarry R, et al. New York: Springer New York; 2005:397–420.
163. Altman N: **Differential Expression Analysis using LIMMA**. https://www.bioconductor.org/packages/release/data/experiment/vignettes/ChimpHumanBrainData/inst/doc/DiffExpressVignette.pdf. Bioconductor project. 2016-11-02.
163. Smyth GK: **Estrogen Data A 2x2 factorial experiment**. https://www.bioconductor.org/help/course-materials/2005/BioC2005/labs/lab01/estrogen/. Bioconductor project. 2016-11-02.
164. Smyth GK, Hu Y, Ritchie M, Silver J, et al.: **limma**. http://bioconductor.org/packages/release/bioc/manuals/limma/man/limma.pdf. Bioconductor project. 2016-07-11.
165. Smyth GK, Michaud J, Scott HS: **Use of within-array replicate spots for assessing differential expression in microarray experiments**. *Bioinformatics* 2005, **21**:2067–2075.
166. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society, Series B* 1995, **57**:289–300.
167. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95**:14863–14868.
168. Wilkinson L, Friendly M: **The History of the Cluster Heat Map**. *The American Statistician* 2009, **63**:179–184.
169. Shannon W, Culverhouse R, Duncan J: **Analyzing microarray data using cluster analysis**. *Pharmacogenomics* 2003, **4**:41–52.
170. Gehlenborg N, Wong B: **Points of view: Heat maps**. *Nature Methods* 2012, **9**:213.
171. Datta S, Datta S: **Comparisons and validation of statistical clustering techniques for microarray gene expression data**. *Bioinformatics* 2003, **19**:459–466.
172. D'haeseleer P: **How does gene expression clustering work?** *Nature Biotechnology* 2005, **23**:1499–1501.
173. Girke T: **R & Bioconductor Manual**. http://manuals.bioinformatics.ucr.edu/home/R_BioCondManual#TOC-Hierarchical-Clustering-HC-. UC Riverside. 2016-11-28.
174. R Core Team: **The R Stats Package**. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/stats-package.html. R Core Team. 2016-11-28.
175. Warnes G R, Bolker B, Bonebakker L, Gentleman R, et al.: **Package 'gplots'**. https://cran.r-project.org/web/packages/gplots/gplots.pdf. CRAN. 2016-11-28.
176. Neuwirth, E.: **Package 'RColorBrewer'**. https://cran.r-project.org/web/packages/RColorBrewer/RColorBrewer.pdf. CRAN. 2016-11-28.
177. R Core Team: **The R Graphics Package**. https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/graphics-package.html. R Core Team. 2016-11-28.
178. Huang DW, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists**. *Nucleic Acids Research* 2009, **37**:1–13.
179. Ashburner M, Ball CA, Blake JA, Botstein D, et al.: **Gene Ontology: tool for the unification of biology**. *Nature Genetics* 2000, **25**:25–29.
180. Tipney H, Hunter L: **An introduction to effective use of enrichment analysis software**. *Humane Genomics* 2010, **4**:1–5.
181. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nature Protocols* 2008, **4**:44–57.

182. Khatri P, Draghici S, Ostermeier GC, Krawetz SA: **Profiling Gene Expression Using Onto-Express**. *Genomics* 2002, **79**:266–270.
183. Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, et al.: **MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data**. *Genome Biology* 2003, **4**:R7.
184. Zeeberg BR, Feng W, Wang G, Wang MD, et al.: **GoMiner: a resource for biological interpretation of genomic and proteomic data**. *Genome Biology* 2003, **4**:R28.
185. Dennis G, Shermann BT, Hosack DA, Yang J, et al.: **DAVID: Database for Annotation, Visualization, and Integrated Discovery**. *Genome Biology* 2003, **4**:P3.
186. Hosack DA, Dennis G, Sherman BT, Lane HC, et al.: **Identifying biological themes within lists of genes with EASE**. *Genome Biology* 2003, **4**:R70.
187. Martínez-Cruz LA, Rubio A, Martínez-Chantar ML, Labarga A, et al.: **GARBAN: genomic analysis and rapid biological annotation of cDNA microarray and proteomic data**. *Bioinformatics* 2003, **19**:2158–2160.
188. Castillo-Davis CI, Hartl DL: **GeneMerge—post-genomic analysis, data mining, and hypothesis testing**. *Bioinformatics* 2003, **19**:891–892.
189. Berriz GF, King OD, Bryant B, Sander C, et al.: **Characterizing gene sets with FuncAssociate**. *Bioinformatics* 2003, **19**:2502–2504.
190. Curtis RK, Orešič M, Vidal-Puig A: **Pathways to the analysis of microarray data**. *Trends in Biotechnology* 2016, **23**:429–435.
191. Khatri P, Drăghici S: **Ontological analysis of gene expression data: current tools, limitations, and open problems**. *Bioinformatics* 2005, **21**:3587–3595.
192. Hung J-H, Yang T-H, Hu Z, Weng Z, et al.: **Gene set enrichment analysis: performance evaluation and usage guidelines**. *Briefings in Bioinformatics* 2012, **13**:281–291.
193. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, et al.: **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences* 2005, **102**:15545–15550.
194. Tuimala J: **DNA microarray data analysis using Bioconductor.** http://tuimala.mbnet.fi/oppaat/r2.pdf. CSC, the Finnish IT center for Science. 2016-11-28.
195. Gentleman R, et al.: **Package 'Category'**. https://www.bioconductor.org/packages/release/bioc/manuals/Category/man/Category.pdf. Bioconductor project. 2016-12-5.
196. Gentleman R, Falcon S.: **Package 'GOstats'**. https://bioconductor.org/packages/release/bioc/manuals/GOstats/man/GOstats.pdf. Bioconductor project. 2016-12-5.
197. A S, A B-H: **Hierarchical classification of gene ontology terms using the GOstruct method**. *Journal of Bioinformatics and Computational Biology* 2010, **8**:357–376.
198. Cheng L, Lin H, Hu Y, Wang J, Yang Z: **Gene Function Prediction Based on the Gene Ontology Hierarchical Structure**. *PLoS One* 2014, **9**:e107187.
199. Rivals I, Personnaz L, Taing L, Potier M-C: **Enrichment or depletion of a GO category within a class of genes: which test?** *Bioinformatics* 2006, **23**:401–407.
200. Beißbarth T, Speed TP: **GOstat: find statistically overrepresented Gene Ontologies within a group of genes**. *Bioinformatics* 2004, **20**:1464–1465.
201. Yu G, Petyuk VA: **Package 'ReactomePA'**. http://bioconductor.org/packages/release/bioc/manuals/ReactomePA/man/ReactomePA.pdf. Bioconductor project. 2017-24-3.
202. Hotelling H: **Analysis of a complex of statistical variables into principal components**. *Journal of Educational Psychology* 1933, **24**:417–441.
203. Chessel D, Dufour AB, Thioulouse J: **The ade4 package - I : One-table methods**. *R News* 2004, **4**:5–10.
204. Dray S, Dufour A-B: **The ade4 Package: Implementing the Duality Diagram for Ecologists**. *Journal of Statistical Software* 2007, **22**:1–20.
205. Lê S, Josse J, Husson F: **FactoMineR: An R Package for Multivariate Analysis**. *Journal of Statistical Software* 2008, **25**:1–18.
206. Nenadic O, Greenacre M: **Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package**. *Journal of Statistical Software* 2007, **20**:1–13.
207. Everitt BS, Dunn G: **Principal Components Analysis**. In *Applied Multivariate Data Analysis*. Edited by Ueberberg A. West Sussex, United Kingdom: John Wiley & Sons; 2001:48–73.
208. Ringner M: **What is principal component analysis?** *Nature Biotechnology* 2008, **26**:303–304.
209. Hilsenbeck SG, William E, Schiff R, Connell O, et al.: **Statistical Analysis of Array Expression Data as Applied to the Problem of Tamoxifen Resistance**. *Journal of the National Cancer Institute* 1999, **91**:453–459.
210. Vohradsky J, Li XM, Thomson CJ: **Identification of procaryotic developmental stages by statistical analyses of two-dimensional gel patterns**. *Electrophoresis* 1997, **18**:1418–1428.
211. Lee S-I, Batzoglou S: **Application of independent component analysis to microarrays**. *Genome Biology* 2003, **4**:R76.

212. Frigyesi A, Veerla S, Lindgren D, Höglund M: **Independent component analysis reveals new and biologically significant structures in micro array data**. *BMC Bioinformatics* 2006, **7**:290.
213. Abdi H, Williams LJ: **Principal component analysis**. *Wiley Interdisciplinary Reviews: Computational Statistics* 2010, **2**:433–459.
214. Fodor IK. **A survey of dimension reduction techniques**. Center for Applied Scientific Computing, Lawrence Livermore National Laboratory. Livermore, CA, United Sates. 2002.
215. Alter O, Brown PO, Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling**. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**:10101–10106.
216. Wall ME, Rechensteiner A, Rocha LM: **Singular value decomposition and principal component analysis**. In *A Practical Approach to Microarray Data Analysis*. Edited by Berrar DP, Dubitzky W, Granzow M. New York, Boston, Dordrecht, London, Moscow: Kluwer Academic Publisher; 2003:91–109.
217. Alberti G: **An R Script to facilitate correspondence analysis. A guide to the use and the interpretation of results from an archaeological perspective**. *Archeologia e Calcolatori* 2013, **24**:25–53.
218. Alberti G: **CAinterprTools: An R package to help interpreting Correspondence Analysis' results**. *SoftwareX* 2015, **1–2**:26–31.
219. Dray S, Dufour AB, Thioulouse J, Jombart, T, et al.: **Package 'ade4'**. https://cran.r-project.org/web/packages/ade4/ade4.pdf. CRAN. 2017-24-3.
220. Husson F, Josse J, Le S, Mazet J: **Package 'FactoMineR'**. https://cran.r-project.org/web/packages/FactoMineR/FactoMineR.pdf. CRAN. 2017-24-3.
221. Greenacre M, Nenadic O, Friendly M: **Package 'ca'**. https://cran.r-project.org/web/packages/ca/ca.pdf. CRAN. 2017-24-3.
222. Kendall MG: *A Course in Multivariate Analysis*. 5 edn. London: Griffin; 1957.
223. Curran PJ, Hussong AM: **Integrative Data Analysis: The Simultaneous Analysis of Multiple Data Sets**. *Psychological Methods* 2009, **14**:81–100.
224. Culhane AC, Thioulouse J, Perrière G, Higgins DG: **MADE4: an R package for multivariate analysis of gene expression data**. *Bioinformatics* 2005, **21**:2789–2790.
225. Culhane A: **Integrated pathway analysis of multiple 'omics datasets**. https://www.bioconductor.org/help/course-materials/2014/BioC2014/IntAnalysis.html. Bioconductor project. 2017-2-4.
226. R Development Core Team . **R: A language and environment for statistical computing**. http://www.R-project.org/. R Foundation for Statistical Computing, Vienna, Austria. 2016.
227. The Bioconductor Team: **Package 'hgu95acdf.db'**. https://bioconductor.org/packages/release/data/annotation/manuals/hgu95acdf/man/hgu95acdf.pdf. Bioconductor project. 2016-07-11.
228. The Bioconductor Team: **Package 'hgu95bcdf.db'**. https://bioconductor.org/packages/release/data/annotation/manuals/hgu95bcdf/man/hgu95bcdf.pdf. Bioconductor project. 2016-07-11.
229. The Bioconductor Team: **Package 'hgu95ccdf.db'**. https://bioconductor.org/packages/release/data/annotation/manuals/hgu95ccdf/man/hgu95ccdf.pdf. Bioconductor project. 2016-07-11.
230. The Bioconductor Team: **Package 'hgu95dcdf.db'**. https://bioconductor.org/packages/release/data/annotation/manuals/hgu95dcdf/man/hgu95dcdf.pdf. Bioconductor project. 2016-07-11.
231. The Bioconductor Team: **Package 'hgu95ecdf.db'**. https://bioconductor.org/packages/release/data/annotation/manuals/hgu95ecdf/man/hgu95ecdf.pdf. Bioconductor project. 2016-07-11.
232. The Bioconductor Team: **Package 'hgu133acdf'**. https://bioconductor.org/packages/release/data/annotation/manuals/hgu133acdf/man/hgu133acdf.pdf. Bioconductor project. 2016-07-11.
233. The Bioconductor Team: **Package** 'hgu133bcdf'. https://bioconductor.org/packages/release/data/annotation/manuals/hgu133bcdf/man/hgu133bcdf.pdf. Bioconductor project. 2016-07-11.
234. The Bioconductor Team: **Package 'hgu133plus2cdf'**. https://bioconductor.org/packages/release/data/annotation/manuals/hgu133plus2cdf/man/hgu133plus2cdf.pdf. Bioconductor project. 2016-07-11.
235. Walker A, Braglia L: **Package 'openxlsx'**. https://cran.r-project.org/web/packages/openxlsx/openxlsx.pdf. CRAN. 2016-07-11.
236. Gentleman R, Carey V, Morgan M, Falcon S: **Package 'Biobase'**. http://bioconductor.org/packages/release/bioc/manuals/Biobase/man/Biobase.pdf. Bioconductor project. 2016-07-11.
237. The Bioconductor Dev Team: **Package 'BiocGenerics'**. https://bioconductor.org/packages/release/bioc/manuals/BiocGenerics/man/BiocGenerics.pdf. Bioconductor

project. 2016-07-11.
238. R-core Team: **Package 'parallel'**. https://stat.ethz.ch/R-manual/R-devel/library/parallel/doc/parallel.pdf. R-core. 2016-07-11.
239. Wickham H, Müller K: **Package 'DBI'**. https://cran.r-project.org/web/packages/DBI/DBI.pdf. CRAN. 2016-07-11.
240. Pages H, Carlson M, Falcon S, Li N: **Package 'AnnotationDbi'**. https://bioconductor.org/packages/release/bioc/manuals/AnnotationDbi/man/AnnotationDbi.pdf. Bioconductor project. 2016-07-11.
241. Langfelder P, Zhang B, Horvath S: **Package 'dynamicTreeCut'**. https://cran.r project.org/web/packages/dynamicTreeCut/dynamicTreeCut.pdf. CRAN. 2016-07-11.
242. Müllner D: **Package 'fastcluster'**. https://cran.r-project.org/web/packages/fastcluster/fastcluster.pdf. CRAN. 2016-07-11.
243. Langfelder P, Horvath S, Cai C, Dong J, et al.: **Package 'WGCNA'**. https://cran.r-project.org/web/packages/WGCNA/WGCNA.pdf. CRAN. 2016-07-11.
244. R-core Team: **Package 'stats4'**. https://stat.ethz.ch/R-manual/R-devel/library/stats4/html/stats4-package.html. R-core. 2016-07-11.
245. Pagés H, Aboyoun P, Lawrence M: **Package 'IRanges'**. https://www.bioconductor.org/packages/devel/bioc/manuals/IRanges/man/IRanges.pdf. Bioconductor project. 2016-07-11.
246. Pagès H, Lawrence M, Aboyoun P: **Package 'S4Vectors'**. https://www.bioconductor.org/packages/release/bioc/manuals/S4Vectors/man/S4Vectors.pdf. Bioconductor project. 2016-07-11.
247. Wickham H, James D A, Falcon S, et al.: **Package 'RSQLite'**. https://cran.r-project.org/web/packages/RSQLite/RSQLite.pdf . CRAN. 2016-07-11.
248. Carlson M: **Package 'org.Hs.eg.db'**. https://bioconductor.org/packages/release/data/annotation/manuals/org.Hs.eg.db/man/org.Hs.eg.db.pdf. Bioconductor project. 2016-07-11.
249. Gentleman R: **Package 'annotate'**. https://www.bioconductor.org/packages/release/bioc/manuals/annotate/man/annotate.pdf. Bioconductor project. 2016-07-11.
250. Temple Lang D: **Package 'XML'**. https://cran.r-project.org/web/packages/XML/XML.pdf. CRAN. 2016-07-11.
251. Carlson M: **Package 'GO.db'**. https://bioconductor.org/packages/release/data/annotation/manuals/GO.db/man/GO.db.pdf. Bioconductor project. 2016-07-11.
252.Tenenbaum D, et al.: **Package 'BiocInstaller'**. https://bioconductor.org/packages/release/bioc/manuals/BiocInstaller/man/BiocInstaller.pdf. Bioconductor project. 2016-07-11.
253. Grothendieck G: **Package 'sqldf'**. https://cran.r-project.org/web/packages/sqldf/sqldf.pdf. CRAN. 2016-07-11.
254. Chen H: **Package 'VennDiagram'**. https://cran.r-project.org/web/packages/VennDiagram/VennDiagram.pdf. CRAN. 2016-07-11.
255. R-core Team: **Package 'graphics'**. https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/graphics-package.html. R-core. 2016-07-11.
256. Dowle M, Srinivasan A, Short T, Lianoglo S, et al.: **Package 'data.table'**. https://cran.r-project.org/web/packages/data.table/data.table.pdf.CRAN. 2016-07-11.
257. Wickham H, et al.: **Package 'scales'**. https://cran.r-project.org/web/packages/scales/scales.pdf. CRAN. 2016-07-11.
258. Wickham H, Francois R, Henry L, Müller K: **Package 'dplyr'**. https://cran.r-project.org/web/packages/dplyr/dplyr.pdf. CRAN. 2016-07-11.
259. Irizarry RA, Love MI : **Package 'rafalib'**. https://cran.r-project.org/web/packages/rafalib/rafalib.pdf. CRAN. 2016-07-11.
260. Kassambara A, Mundt F: **Package 'factoextra'**. https://cran.r-project.org/web/packages/factoextra/factoextra.pdf. CRAN. 2016-07-11.
261. Lander ES, Linton LM, Birren B, Nusbaum C, et al.: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860–921.
262. Noble I: **Human genome finally complete.** http://news.bbc.co.uk/2/hi/science/nature/2940601.stm. BBC News. 2003-04-14.
263. Kolata G: **Human Genome, Then and Now**. http://www.nytimes.com/2013/04/16/science/the-human-genome-project-then-and-now.html. The New York Times. 2013-04-15.
264. National Center for Biotechnology Information: **Introducing the New Human Genome Assembly: GRCh38**. https://ncbiinsights.ncbi.nlm.nih.gov/2013/12/24/introducing-the-new-human-genome-assembly-grch38/. NCBI Insights. 2013-12-24.

265. Huntley RP, Sawford T, Martin MJ, O'Donovan C: **Understanding how and why the Gene Ontology and its annotations evolve: the GO within UniProt**. *Gigascience* 2014, **3**:4.
266. Wadi L, Meyer M, Weiser J, Stein LD, et al.: **Impact of outdated gene annotations on pathway enrichment analysis**. *Nature Methods* 2016, **13**:705–706.
267. Emmert-Streib F, de Matos Simoes R, Mullan P, Haibe-Kains B, Dehmer M: **The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks**. *Frontiers in Genetics* 2014, **5**:15.
268. Chen D, Yang H: **Integrated analysis of differentially expressed genes in breast cancer pathogenesis**. *Oncology Letters* 2015, **9**:2560–2566.
269. Liu H, D'Andrade P, Fulmer-Smentek S, Lorenzi P, et al.: **mRNA and microRNA expression profiles of the NCI-60 integrated with drug activities**. *Molecular Cancer Therapeutics* 2010, **9**:1080–1091.
270. Weinstein JN: **Spotlight on molecular profiling: "Integromic" analysis of the NCI-60 cancer cell lines**. *Molecular Cancer Therapeutics* 2006, **5**:2601–2605.
271. Bussey KJ, Chin K, Lababidi S, Reimers M, et al.: **Spotlight on molecular profiling: "Integromic" analysis of the NCI-60 cancer cell lines**. *Molecular Cancer Therapeutics* 2006, **5**:853–867.
272. Scherf U, Ross D, Waltham M, Smith L, et al.: **A gene expression database for the molecular pharmacology of cancer**. *Nature Genetics* 2000, **24**:236–244.
273. Weinstein JN: **Integromic analysis of the NCI-60 cancer cell lines**. *Breast Disease* 2004, **19**:11–22.
274. Blower PE, Verducci JS, Lin S, Zhou J, et al.: **MicroRNA expression profiles for the NCI-60 cancer cell panel**. *Molecular Cancer Therapeutics* 2007, **6**:1483–1491.
275. Gaur A, Jewell DA, Liang Y, Ridzon D, et al.: **Characterization of MicroRNA Expression Levels and Their Biological Correlates in Human Cancer Cell Lines**. *Cancer Research* 2007, **67**:2456–2468.
276. Perou CM, Sørlie T, Eisen MB, van de Rijn M, et al.: **Molecular portraits of human breast tumours**. *Nature* 2000, **406**:747–752.
277. Stinson SF, Alley MC, Kopp WC, Fiebig HH, et al.: **Morphological and immunocytochemical characteristics of human tumor cell lines for use in a disease-oriented anticancer drug screen**. *Anticancer Research* 1992, **12**:1035–1053.
278. Fredman P, Hedberg K, Brezicka T, et al.: **Gangliosides as therapeutic targets for cancer**. *BioDrugs* 2003, **17**:155–167.
279. Stuelten CH, Mertins SD, Busch JI, Gowens M, et al.: **Complex display of putative tumor stem cell markers in the NCI60 tumor cell line panel**. *Stem Cells* 2010, **28**:649–660.
280. Chai J: **Validity of Markers for Epithelial Cells and Mesenchymal Cells**. In *Stem Cells and Cancer Stem Cells,Volume 10*. Edited by Hayat MA. Dordrecht: Springer Netherlands; 2013:23–33.
281. Sudo T, Iwaya T, Nishida N, Sawada G, et al.: **Expression of Mesenchymal Markers Vimentin and Fibronectin: The Clinical Significance in Esophageal Squamous Cell Carcinoma**. *Annals of Surgical Oncology* 2013, **20**:324–335.
282. Ramsay RG, Gonda T: **MYB function in normal and cancer cells**. *Nature Reviews Cancer* 2008, **8**:523–534.
283. Liu G, Yang J, Zhao Y, Wang Z, et al.: **Expression of secretory leukocyte protease inhibitor detected by immunohistochemistry correlating with prognosis and metastasis in colorectal cancer**. *World Journal of Surgical Oncology* 2014, **12**:369.
284. Ishii H, Saitoh M, Sakamoto K, Kondo T, et al.: **Epithelial Splicing Regulatory Proteins 1 (ESRP1) and 2 (ESRP2) Suppress Cancer Cell Motility via Different Mechanisms**. *The Journal of Biological Chemistry* 2014, **289**:27386–27399.
285. Leontieva O V, Ionov Y: **RNA-binding motif protein 35A is a novel tumor suppressor for colorectal cancer**. *Cell Cycle* 2009, **8**:490–497.
286. Rezaul K, Thumar JK, Lundgren DH, Eng JK, et al.: **Differential Protein Expression Profiles in Estrogen Receptor–Positive and –Negative Breast Cancer Tissues Using Label-Free Quantitative Proteomics**. *Genes & Cancer* 2010, **1**:251–271.
287. Parkkila S, Pan P, Ward A, Gibadulinova A, et al.: **The calcium-binding protein S100P in normal and malignant human tissues**. *BMC Clinical Pathology* 2008, **8**:2.
288. Arumugam T, Logsdon CD: **S100P: a novel therapeutic target for cancer**. *Amino Acids* 2011, **41**:893–899.
289. Reemann P, Reimann E, Ilmjärv S, Porosaar O, et al.: **Melanocytes in the Skin – Comparative Whole Transcriptome Analysis of Main Skin Cell Types**. *PLoS One* 2014, **9**:e115717.
290. Nissan X, Larribere L, Saidani M, Hurbain I, et al.: **Functional melanocytes derived from human pluripotent stem cells engraft into pluristratified epidermis**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**:14861–14866.
291. Haq R, Yokoyama S, Hawryluk EB, Jönsson GB, et al.: **BCL2A1 is a lineage-specific antiapoptotic melanoma oncogene that confers resistance to BRAF inhibition**. *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**:4321–4326.
292. Vogler M: **BCL2A1: the underdog in the BCL2 family**. *Cell Death and Differentiation* 2012, **19**:67–74.
293. Agarwal R, Jurisica I, Mills GB, Wa Cheng K: **The Emerging Role of the RAB25 Small GTPase in Cancer**. *Traffic* 2009, **10**:1561–1568.
294. Mitra S, Cheng KW, Mills GB: **Rab25 in Cancer: A brief update**. *Biochemical Society Transactions* 2012, **40**:1404–1408.

295. Rhee I, Veillette A: **Protein tyrosine phosphatases in lymphocyte activation and autoimmunity**. *Nature immunology* 2012, **13**:439–447.
296. Joshi AD, Hegde G V, Dickinson JD, Mittal AK, et al.: **ATM, CTLA4, MNDA, and HEM1 in High versus Low CD38–Expressing B-Cell Chronic Lymphocytic Leukemia**. *Clinical Cancer Research* 2007, **13**:5295–5304.
297. Wu M, Hamaker M, Li L, Small D, et al.: **DOCK2 Interacts With FLT3 and Modulates the Survival of FLT3-Expressing Leukemia Cells**. *Leukemia* 2016, **31**:688–696.
298. Shima K, Nosho K, Baba Y, Cantor M, et al.: **Prognostic Significance of CDKN2A (p16) Promoter Methylation and Loss of Expression in 902 Colorectal Cancers: Cohort Study and Literature Review**. *International Journal of Cancer* 2011, **128**:1080–1094.
299. Xing X, Cai W, Shi H, Wang Y, et al.: **The prognostic value of CDKN2A hypermethylation in colorectal cancer: a meta-analysis**. *British Journal of Cancer* 2013, **108**:2542–2548.
300. Mohamed A, Gonzalez R, Lawson D, Wang J, et al.: **SOX10 expression in malignant melanoma, carcinoma, and normal tissues**. *Applied Immunohistochemistry & Molecular Morphology* 2013, **21**:506–510.
301. Fuchs I, Lichtenegger W, Buehler H, Henrich W, et al.: **The prognostic significance of epithelial-mesenchymal transition in breast cancer**. *Anticancer Research* 2002, **22**:3415–3419.
302. Taetle R, Ostergaard H, Smedsrud M, Trowbridge I: **Regulation of CD45 expression in human leukemia cells**. *Leukemia* 1991, **5**:309–314.
303. Yamanashi Y, Okada M, Semba T, Yamori T, et al.: **Identification of HS1 protein as a major substrate of protein-tyrosine kinase(s) upon B-cell antigen receptor-mediated signaling.** *Proceedings of the National Academy of Sciences of the United States of America* 1993, **90**:3631–3635.
304. Gelse K, Pöschl E, Aigner T: **Collagens--structure, function, and biosynthesis**. *Advanced Drug Delivery Reviews* 203AD, **55**:1531–1546.
305. Medici D, Nawshad A: **Type I collagen promotes epithelial-mesenchymal transition through ILK-dependent activation of NF-kappaB and LEF-1**. *Matrix Biology* 2010, **29**:161–165.
306. Shintani Y, Maeda M, Chaika N, Johnson KR, et al.: **Collagen I promotes epithelial-to-mesenchymal transition in lung cancer cells via transforming growth factor-beta signaling**. *American Journal of Respiratory Cell and Molecular Biology* 2008, **38**:95–104.
307. Thiery JP: **Epithelial-mesenchymal transitions in tumour progression**. *Nature Reviews Cancer* 2002, **2**:442–454.
308. Micalizzi DS, Farabaugh SM, Ford HL: **Epithelial-mesenchymal transition in cancer: parallels between normal development and tumor progression**. *Journal of Mammary Gland Biology and Neoplasia* 2010, **15**:117–134.
309. D S, Jiang K, Fu Y, Fang R, et al.: **Overexpression of SPARC correlates with poor prognosis in patients with cervical carcinoma and regulates cancer cell epithelial-mesenchymal transition**. *Oncology Letters* 2016, **11**:3251–3258.
310. National Center for Biotechnology Information: **SPARC secreted protein acidic and cysteine rich [*Homo sapiens* (human)]**. https://www.ncbi.nlm.nih.gov/gene/6678. 2017-07-02.
311. Meng C, Culhane A: **Integrative Exploratory Analysis of Two or More Genomic Datasets**. *Methods in Molecular Biology* 2016, **1418**:19–38.
312. Zippelius A, Gati A, Bartnick T, Walton S, et al.: **Melanocyte differentiation antigen RAB38/NY-MEL-1 induces frequent antibody responses exclusively in melanoma patients**. *Cancer Immunology, Immunotherapy* 2007, **56**:249–258.
313. Cool HEM, Baxter MJ: **Cemeteries and significance tests**. *Journal of Roman Archaeology* 2005, **18**:397-403.

6 Appendix

6.1 Differential expression analysis

Supplementary Table 1: DEA of the gene expression data of the second data set from the HG-U133 plus 2.0 platform with absolute log2-fold change 1 in the upper triangular matrix and DEA of the protein expression data of the second data set from the proteome data with absolute log2-fold change 1 in the lower triangular matrix

	BR	CNS	CO	LU	LE	ME	OV	PR	RE
BR		0; 1	20;16	0; 0	83; 33	57; 123	0; 0	0; 5	1; 6
CNS	2; 5		618; 623	13; 24	647; 319	338; 351	59; 59	59; 130	64; 99
CO	1; 2	49; 42		60; 66	353; 130	679; 682	14;19	10; 50	502; 551
LU	0; 0	1; 0	10; 3		372; 185	183; 245	0; 0	1; 10	0; 1
LE	15; 8	140; 37	65; 21	68; 26		462; 790	90; 243	19; 146	579; 822
ME	0; 0	13; 6	56; 26	2; 0	32; 99		276; 248	83; 220	647; 631
OV	0; 1	2; 2	0; 1	0; 0	29; 64	1; 2		1; 9	1; 5
PR	3; 63	17; 63	13; 31	2; 26	9; 126	3; 35	5; 25		54; 15
RE	0; 3	6; 9	19; 23	1; 1	34; 85	17; 19	0; 1	27; 4	

Supplementary Table 2: DEA of the gene expression data of the third data set from the HG-U133 plus 2.0 platform with absolute log2-fold change 1 in the upper triangular matrix and DEA of the protein expression data of the third data set from the proteome data with absolute log2-fold change 1 in the lower triangular matrix

	BR	CNS	CO	LU	LE	ME	OV	PR	RE
BR		0; 1	7; 8	0; 0	26; 11	17; 50	0; 0	0; 3	0; 2
CNS	1; 4		219; 283	2; 7	207; 99	100; 154	14; 30	26; 62	22; 35
CO	1; 2	43; 38		22; 18	136; 38	285; 245	6 ;2	4; 24	211; 177
LU	0; 0	1; 0	9; 3		139; 45	78; 92	0; 0	1; 3	0; 0
LE	14; 8	124; 36	61; 19	62; 25		149; 280	24; 78	8; 64	195; 273
ME	0; 0	11; 6	53; 25	2; 0	32; 89		110; 98	31; 97	242; 229
OV	0; 1	2; 2	0; 1	0; 0	28; 54	0; 0		1; 5	1; 1
PR	3; 61	16; 60	13; 31	2; 25	9; 118	3; 33	5; 25		28; 6
RE	0; 3	6; 9	16; 23	1; 1	31; 71	17; 17	0; 1	26; 4	

Supplementary Table 3: DEA of the gene expression data of the first data set from the HG-U133 plus 2.0 platform with absolute log2-fold change 1, 1.4 and 2 (in the following order in the table)

Tissue comparison data	Total number of genes – log_2-fold change 1, 1.4 and 2	Number of de genes – over-expressed –log_2-fold change 1, 1.4 and 2	Number of de genes – under-expressed – log_2-fold change 1, 1.4 and 2	Number of unchanged genes – log_2-fold change 1, 1.4 and 2
BR_LE	4,097; 2,250; 1,091	2; 2; 1	7; 7; 7	4,088; 2,241; ,1083
BR_LC	4,968; 2,932; 1,525	46; 42; 36	30; 22; 18	4,892; 2,868; 1,471
BR_ME	6,589; 4,153; 2,274	188; 181; 155	78; 72; 58	6,323; 3,900; 2,061
BR_OV	2,923; 1,362; 500	0; 0; 0	0; 0; 0	2,923; 1,362; 500
BR_PR	5,500; 3,229; 1,714	217; 154; 104	333; 286; 249	4,950; 2,789; 1,361
BR_RE	3,223; 1,691; 774	0; 0; 0	0; 0; 0	3,223; 1,691; 774
CNS_CO	306; 219; 122	1; 0; 0	14; 13; 11	291; 206; 111
CNS_LE	4,029; 2,176; 1,023	4; 4; 1	12; 12; 10	4,013; 2,160; 1,012
CNS_LC	6,378; 4,113; 2,449	1,477; 1,239; 978	1,265; 883; 534	3,636; 1,991; 937
CNS_ME	7,386; 4,773; 2,823	1,545; 1,380; 1,119	803; 518; 281	5,038; 2,875; 1,423
CNS_OV	3,716; 2,079; 1,028	49; 44; 38	42; 36; 32	3,625; 1,999; 958
CNS_PR	5,095; 3,048; 1,657	853; 647; 477	716; 520; 347	3,526; 1,881; 833
CNS_RE	4,393; 2,623; 1,365	148; 137; 117	110; 101; 81	4,135; 2,385; 1,167
CO_LE	6,546; 4,002; 2,195	124; 106; 89	247; 209; 162	6,175; 3,687; 1,944
CO_LC	4,133; 2,306; 1,078	193; 131; 87	245; 206; 160	3,695; 1,969; 831
CO_ME	5,838; 3,715; 2,120	758; 684; 568	321; 267; 204	4,759; 2,764; 1,348
CO_OV	4,373; 2,515; 1,283	145; 102; 69	158; 146; 123	4,070; 2,267; 1,091
CO_PR	6,212; 3,941; 2,278	1,540; 1,068; 684	1,609; 1,310; 968	3,063; 1,563; 626
CO_RE	4,039; 2,352; 1,181	23; 21; 15	45; 42; 39	3,971; 2,289; 1,127
LC_LE	5,542; 3,353; 1,700	14; 9; 6	106; 97; 83	5,422; 3,247; 1,611
LC_ME	5,916; 3,655; 2,054	1,078; 674; 367	1,389; 1,203; 933	3,449; 1,778; 754
LC_OV	6,784; 4,171; 2,309	843; 747; 598	552; 387; 219	5,389; 3,037; 1,492
LC_PR	4,692; 2,844; 1,520	411; 317; 214	563; 481; 378	3,718; 2,046; 928
LC_RE	2,469; 1,204; 462	0; 0; 0	0; 0; 0	2,469; 1,204; 462
LE_ME	2,056; 1,202; 592	2; 1; 1	23; 18; 12	2,031; 1,183; 579
LE_OV	3,165; 1,642; 692	0; 0; 0	8; 8; 7	3,157; 1,634; 685
LE_PR	7,211; 4,538; 2,589	1,347; 858; 428	1,863; 1,568; 1,232	4,001; 2,112; 929
LE_RE	6,516; 4,131; 2,325	240; 199; 148	577; 547; 468	5,699; 3,385; 1,709
ME_OV	7,744; 5,061; 2,973	41; 39;36	289; 274; 248	7,414; 4,748; 2,689
ME_PR	7,934; 5,193; 2,935	1,455; 961; 475	1,984; 1,795; 1,445	4,495; 2,437; 1,015
ME_RE	5,349; 3,278; 1,772	656; 555; 427	583; 481; 322	4,110; 2,242; 1,023
OV_PR	7,136; 4,522; 2,442	173; 151; 109	473; 390; 286	6,490; 3,981; 2,047
OV_RE	5,781; 3,545; 2,005	1,369; 938; 599	1,506; 1,162; 796	2,906; 1,445; 610
PR_RE	4,814; 2,654; 1,180	7; 6; 3	20; 19; 18	4,787; 2,629; 1,159
BR_LE	3,681; 1,979; 897	3; 2; 2	16; 16;13	3,662; 1,961; 882
BR_LC	5,113; 3,144; 1,623	128; 98; 74	35; 34; 27	4,950; 3,012; 1,522

Supplementary Table 4: DEA of the gene expression data of the second data set from the HG-U133 plus 2.0 platform with absolute log2-fold change 1, 1.4 and 2 (in the following order in the table)

Tissue comparison data	Total number of genes – log_2-fold change 1, 1.4 and 2	Number of de genes – over-expressed –log_2-fold change 1, 1.4 and 2	Number of de genes – under-expressed – log_2-fold change 1, 1.4 and 2	Number of unchanged genes – log_2-fold change 1, 1.4 and 2
BR_LE	1,923; 1,053; 518	0; 0; 0	1; 1; 1	1,922; 1,052; 517
BR_LC	2,320; 1,369; 735	20; 19; 18	16; 13; 11	2,284; 1,337; 706
BR_ME	3,060; 1,961; 1,073	83; 81; 69	33; 30; 27	2,944; 1,850; 977
BR_OV	1,401; 680; 253	0; 0; 0	0; 0; 0	1,401; 680; 253
BR_PR	2,520; 1,501; 820	57; 43; 31	123; 102; 83	2,340; 1,356; 706
BR_RE	1,541; 839; 373	0; 0; 0	0; 0; 0	1,541; 839; 373
CNS_CO	136; 95; 54	0; 0; 0	5; 4; 4	131; 91; 50
CNS_LE	1847; 1012; 483	1; 1; 0	6; 6; 5	1,840; 1,005; 478
CNS_LC	2,981; 1,935; 1,157	618; 516; 399	623; 439; 285	1,740; 980; 473
CNS_ME	3,391; 2,200; 1,289	647; 588; 469	319; 215; 122	2,425; 1,397; 698
CNS_OV	1,760; 1,000; 494	13; 13; 12	24; 19; 17	1,723; 968; 465
CNS_PR	2,331; 1,395; 764	338; 246; 189	351; 234; 156	1,642; 915; 419
CNS_RE	2,061; 1,228; 657	59; 51; 47	59; 55; 49	1,943; 1,122; 561
CO_LE	3,064; 1,934; 1,078	59; 54; 48	130; 111; 84	2,875; 1,769; 946
CO_LC	1,975; 1,102; 522	64; 39; 28	99; 84; 70	1,812; 979; 424
CO_ME	2,740; 1,784; 1,025	353; 324; 273	130; 113; 85	2,257; 1,347; 667
CO_OV	2,054; 1,202; 611	60; 52; 37	66; 62; 55	1,928; 1,088; 519
CO_PR	2,881; 1,854; 1,104	679; 504; 350	682; 548; 395	1,520; 802; 359
CO_RE	1,932; 1,146; 569	14; 12; 10	19; 17; 16	1,899; 1,117; 543
LC_LE	2,376; 1,490; 822	10; 7; 5	50; 47; 37	2,316; 1,436; 780
LC_ME	2,768; 1,713; 954	502; 329; 184	551; 476; 372	1,715; 908; 398
LC_OV	3,157; 1,953; 1,084	372; 328; 257	185; 136; 83	2,600; 1,489; 744
LC_PR	2,205; 1,343; 723	183; 150; 103	245; 205; 165	1,777; 988; 455
LC_RE	1,196; 584; 215	0; 0; 0	0; 0; 0	1,196; 584; 215
LE_ME	885; 554; 296	1; 0; 0	10; 9; 6	874; 545; 290
LE_OV	1,532; 779; 329	0; 0; 0	1; 1; 1	1,531; 778; 328
LE_PR	3,262; 2,083; 1,210	462; 315; 187	790; 663; 529	2,010; 1,105; 494
LE_RE	3,059; 1,954; 1,096	90; 78; 58	243; 231; 203	2,726; 1,645; 835
ME_OV	3,626; 2,410; 1,436	19; 18; 18	146; 137; 120	3,461; 2,255; 1,298
ME_PR	3,651; 2,390; 1,369	579; 394; 212	822; 756; 614	2,250; 1,240; 543
ME_RE	2,518; 1,552; 842	276; 230; 172	248; 214; 155	1,994; 1,108; 515
OV_PR	3,313; 2,117; 1,185	83; 72; 49	220; 177; 132	3,010; 1,868; 1,004
OV_RE	2,697; 1,654; 938	647; 437; 262	631; 501; 351	1,419; 716; 325
PR_RE	2,259; 1,264; 589	1; 1; 1	9; 8; 8	2,249; 1,255; 580
BR_LE	1,763; 951; 421	1; 0; 0	5; 5; 5	1,757; 946; 416
BR_LC	2,226; 1,415; 752	54; 40; 32	15; 15; 12	2,157; 1,360; 708

Supplementary Table 5: DEA of the gene expression data of the third data set from the HG-U133 plus 2.0 platform with absolute log2-fold change 1, 1.4 and 2 (in the following order in the table)

Tissue comparison data	Total number of genes – log_2-fold change 1, 1.4 and 2	Number of de genes – over-expressed –log_2-fold change 1, 1.4 and 2	Number of de genes – under-expressed – log_2-fold change 1, 1.4 and 2	Number of unchanged genes – log_2-fold change 1, 1.4 and 2
BR_LE	843; 466; 235	0; 0; 0	1; 1; 1	842; 465; 234
BR_LC	1,022; 606; 318	7; 7; 6	8; 6; 6	1,007; 593; 306
BR_ME	1,347; 870; 450	26; 25; 22	11; 10; 10	1,310; 835; 418
BR_OV	643; 314; 123	0; 0; 0	0; 0; 0	643; 314; 123
BR_PR	1,053; 640; 359	17; 13; 9	50; 42; 32	986; 585; 318
BR_RE	656; 383; 177	0; 0; 0	0; 0; 0	656; 383; 177
CNS_CO	54; 39; 25	0; 0; 0	3; 3; 3	51; 36; 22
CNS_LE	813; 450; 219	0; 0; 0	2; 2; 2	811; 448; 217
CNS_LC	1,295; 845; 495	219; 177; 129	283; 197; 132	793; 471; 234
CNS_ME	1,452; 947; 537	207; 186; 151	99; 69; 40	1,146; 692; 346
CNS_OV	783; 441; 214	2; 2; 2	7; 7; 6	774; 432;206
CNS_PR	976; 581; 316	100; 70; 56	154; 101; 68	722; 410; 192
CNS_RE	881; 539; 290	14; 13; 11	30; 28; 26	837; 498; 253
CO_LE	1,325; 862; 485	26; 24; 22	62; 56; 46	1,237; 782; 417
CO_LC	861; 491; 231	22; 12; 6	35; 32; 27	804; 447; 198
CO_ME	1,236; 825; 462	136; 131; 115	38; 36; 25	1,062; 658; 322
CO_OV	910; 533 ; 251	22; 21; 15	18; 16; 14	870; 496; 222
CO_PR	1,239; 814; 503	285; 224; 159	245; 198; 146	709; 392; 198
CO_RE	874; 510; 247	6; 5; 3	2; 2; 2	866; 503; 242
LC_LE	939; 617; 356	4; 3; 3	24; 23; 18	911; 591; 335
LC_ME	1,188; 737; 409	211; 152; 84	177; 152; 119	800; 433; 206
LC_OV	1,388; 860; 479	139; 127; 100	45; 36; 26	1,204; 697; 353
LC_PR	949; 579; 330	78; 64; 48	92; 79; 66	779; 436; 216
LC_RE	563; 283; 124	0; 0; 0	0; 0; 0	563; 283; 124
LE_ME	409; 260; 148	1; 0; 0	3; 2; 2	405; 258; 146
LE_OV	666; 340; 142	0; 0; 0	0; 0; 0	666; 340; 142
LE_PR	1,416; 901; 528	149; 105; 67	280; 237; 196	987; 559; 265
LE_RE	1,335; 883; 463	24; 22; 17	78; 71; 63	1,233; 790; 383
ME_OV	1,585; 1,090; 646	8; 7; 7	64; 61; 53	1,513; 1,022; 586
ME_PR	1,566; 1,024; 584	195; 142; 79	273; 254; 217	1,098; 628; 288
ME_RE	1,069; 671; 384	110; 91; 73	98; 88; 70;	861; 492; 241
OV_PR	1,407; 912; 550	31; 27; 22	97; 79; 61	1,279; 806; 467
OV_RE	1,132; 709; 413	242; 167; 108	229; 183; 141	661; 359; 164
PR_RE	996; 584; 276	1; 1; 1	5; 5; 5	990; 578; 270
BR_LE	795 ; 422; 193	1; 0; 0	1;1; 1	793; 421; 192
BR_LC	940; 595; 324	28; 22; 17	6;6; 5	906; 567; 302

Supplementary Table 6: DEA of the protein expression data of the first data set from the proteome data with absolute log2-fold change 1, 1.4 and 2 (in the following order in the table)

Tissue comparison data	Total number of proteins – log_2-fold change 1, 1.4 and 2	Number of de proteins – over-expressed – log_2-fold change 1, 1.4 and 2	Number of de proteins – under-expressed – log_2-fold change 1, 1.4 and 2	Number of unchanged proteins – log_2-fold change 1, 1.4 and 2
BR_LE	2,524; 1,707; 944	2; 2; 2	5; 5; 5	2,517; 1,700; 937
BR_LC	2,552; 1,711; 986	1; 1; 1	2; 2; 2	2,549; 1,708; 983
BR_ME	2,886; 2,034; 1,267	16; 16; 16	8; 8; 8	2,862; 2,010; 1,243
BR_OV	2,188; 1,332; 649	0; 0; 0	0; 0; 0	2,188; 1,332; 649
BR_PR	2,404; 1,483 ; 792	1; 1; 1	0; 0; 0	2,403; 1,482; 791
BR_RE	2,254; 1,468; 689	0; 0; 0	1;1; 1	2,253; 1,467; 688
CNS_CO	655; 653; 643	4;4;4	65; 65; 65	586; 584; 574
CNS_LE	2,463; 1,536; 799	0; 0; 0	3; 2; 2	2,460; 1,534; 797
CNS_LC	2,535; 1,765; 1,062	55; 52; 51	43;42; 42	2,437; 1,671; 969
CNS_ME	2,735; 2,174; 1,282	150; 148; 148	39; 38; 38	2,546; 1,988; 1,096
CNS_OV	2,273; 1,405; 739	1; 1; 1	0; 0; 0	2,272; 1,404; 738
CNS_PR	2,268; 1,467; 755	15; 15; 15	6; 6; 6	2,247; 1,446; 734
CNS_RE	2,111; 1,549; 671	3; 3;3	2; 2; 2	2,106; 1,544; 666
CO_LE	933; 930; 919	20; 20; 20	63; 63; 63	850; 847; 836
CO_LC	2,194; 1,381; 698	6; 6; 6	9; 9; 9	2,179; 1,366; 683
CO_ME	2,704; 1,859; 1,149	68; 66; 56	21; 19; 19	2,615; 1,774; 1,064
CO_OV	2,232; 1,443; 694	11; 11; 11	3; 3; 3	2,218; 1,429; 680
CO_PR	2,340; 1,601; 875	56; 55; 55	29; 27; 27	2,255; 1,519; 793
CO_RE	2365; 1510; 814	0; 0; 0	1;1; 1	2,364; 1,509; 813
LC_LE	674; 673; 667	13; 13; 13	32; 32; 32	629; 628; 622
LC_ME	2,448; 1,648; 900	21; 21; 21	26; 24; 24	2,401; 1,603; 855
LC_OV	3,043; 2,170; 1,345	77; 73; 73	28; 28; 28	2,938; 2,069; 1,244
LC_PR	2,010; 1,238; 572	2; 2; 2	0;0; 0	2,008; 1,236; 570
LC_RE	2,052; 1,187; 552	0; 0; 0	0; 0; 0	2,052; 1,187; 552
LE_ME	556; 554; 547	2; 2; 2	26; 26; 26	528; 526; 519
LE_OV	2,164; 1,303; 610	1; 1; 1	1; 1; 1	2,162; 1,301; 608
LE_PR	2,918; 2,101; 1,253	33; 33; 33	106; 105; 105	2,779; 1,963; 1,115
LE_RE	2,584; 1,971; 1,076	30; 29; 29	67; 66; 66	2,487; 1,876; 981
ME_OV	1,308; 1,300; 1,291	9; 9; 9	133; 133; 132	1,166; 1,158; 1,150
ME_PR	2,767; 1,943; 1,195	35; 33; 33	99; 96; 93	2,633; 1,814; 1,069
ME_RE	2,206; 1,369; 687	0; 0; 0	1; 1; 1	2,205; 1,368; 686
OV_PR	727; 725; 717	3; 3; 3	36; 36; 36	688; 686; 678
OV_RE	2,320; 1,490; 765	17; 17; 17	20; 20; 20	2,283; 1,453; 728
PR_RE	529; 524; 516	5; 5; 5	27; 27; 27	497; 492; 484
BR_LE	2,085; 1,217; 591	0; 0; 0	1; 1; 1	2,084; 1,216; 590
BR_LC	591; 589; 581	27; 27; 27	4; 4; 4	560; 558; 550

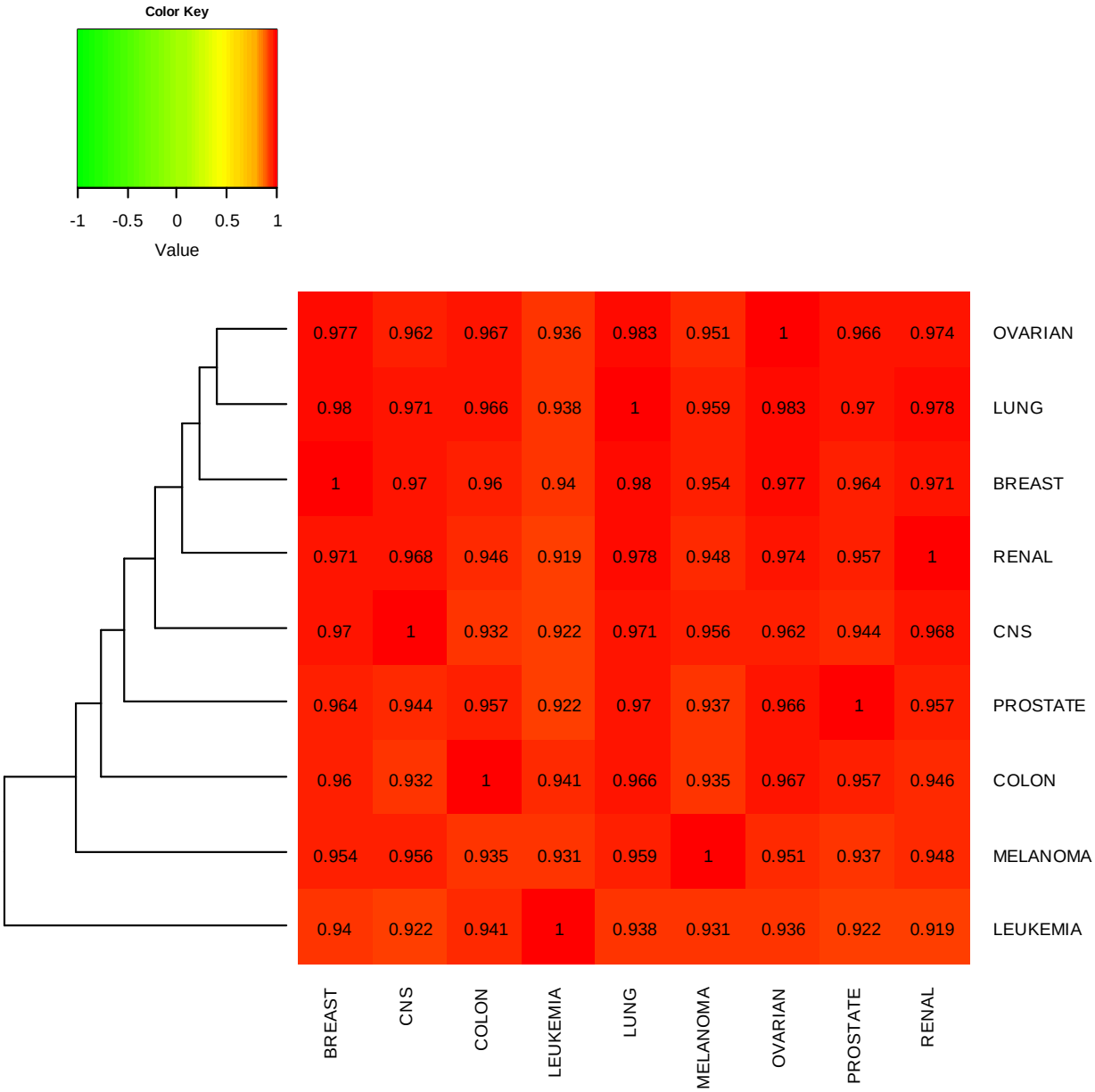
Supplementary Table 7: DEA of the protein expression data of the second data set from the proteome data with absolute log2-fold change 1, 1.4 and 2 (in the following order in the table)

Tissue comparison data	Total number of proteins – log_2-fold change 1, 1.4 and 2	Number of de proteins – over-expressed –log_2-fold change 1, 1.4 and 2	Number of de proteins –under-expressed –log_2-fold change 1, 1.4 and 2	Number of unchanged proteins – log_2-fold change 1, 1.4 and 2
BR_LE	2,394; 1,617; 892	2; 2; 2	5; 5; 5	2,387; 1,610; 885
BR_LC	2,414; 1,628; 943	1; 1; 1	2; 2; 2	2,411; 1,625; 940
BR_ME	2,746; 1,939; 1,209	15; 15; 15	8; 8; 8	2,723; 1,916; 1,186
BR_OV	2,119; 1,265; 613	0; 0; 0	0; 0; 0	2,119; 1,265; 613
BR_PR	2,296; 1,421; 760	0; 0; 0	0 ; 0; 0	2,296; 1,421; 760
BR_RE	2,168; 1,401; 661	0; 0; 0	1; 1; 1	2,167; 1,400; 660
CNS_CO	603; 602; 595	3; 3; 3	63; 63; 63	537; 536; 529
CNS_LE	2,341; 1,461; 768	0; 0; 0	3; 2; 2	2,338; 1,459; 766
CNS_LC	2,409; 1,684; 1,017	49; 48; 48	42; 41; 41	2,318; 1,595; 928
CNS_ME	2,585; 2,046; 1,205	140; 138; 138	37; 36; 36	2,408; 1,872; 1,031
CNS_OV	2,166; 1,338; 706	1; 1; 1	0; 0; 0	2,165; 1,337; 705
CNS_PR	2,164; 1,406; 723	13; 13; 13	6; 6; 6	2,145; 1,387; 704
CNS_RE	1,998; 1,467; 637	2; 2; 2	2; 2; 2	1,994; 1,463; 633
CO_LE	888; 886; 878	17; 17; 17	63; 63; 63	808; 806; 798
CO_LC	2,104; 1,317; 667	6; 6; 6	9; 9; 9	2,089; 1,302; 652
CO_ME	2,583; 1,768; 1,087	65; 63; 63	21; 19; 19	2,497; 1,686; 1,005
CO_OV	2,114; 1,365; 663	10; 10; 10	3; 3; 3	2,101; 1,352; 650
CO_PR	2,227; 1,520; 829	56; 55; 55	26; 25; 25	2,145; 1,440; 749
CO_RE	2,242; 1,430; 775	0; 0; 0	1; 1; 1	2,241; 1,429; 774
LC_LE	636; 636; 632	13; 13; 13	31; 31; 31	592; 592; 588
LC_ME	2,329; 1,569; 855	19; 19; 19	23; 23; 23	2,287; 1,527; 813
LC_OV	2,895; 2,055; 1,268	68; 64; 64	26; 26; 26	2,801; 1,965; 1,178
LC_PR	1,915; 1,182; 544	2; 2; 2	0; 0; 0	1,913; 1,180; 542
LC_RE	1,947; 1,124; 528	0; 0; 0	0; 0; 0	1,947; 1,124; 528
LE_ME	517; 515; 510	2; 2; 2	26; 26; 26	489; 487; 482
LE_OV	2,068; 1,234; 581	1; 1; 1	1; 1; 1	2,066; 1,232; 579
LE_PR	2,786; 2,013; 1,197	32; 32; 32	99; 98; 98	2,655; 1,883; 1,067
LE_RE	2,446; 1,862; 1,014	29; 28; 28	64; 63; 63	2,353; 1,771; 923
ME_OV	2,924; 2,610; 1,816	9; 9; 9	126; 126; 125	2,789; 2,475; 1,682
ME_PR	2,616; 1,833; 1,119	34; 32; 32	85; 83; 80	2,497; 1,718; 1,007
ME_RE	2,103; 1,312; 656	1; 1; 1	2; 2; 2	2,100; 1,309; 653
OV_PR	680; 679; 674	3; 3; 3	35; 35; 35	642; 641; 636
OV_RE	2,214; 1,427; 735	17; 17; 17	19; 19; 19	2,178; 1,391; 699
PR_RE	491; 487; 482	5; 5; 5	25; 25; 25	461; 457; 452
BR_LE	1,985; 1,148; 563	0; 0; 0	1; 1; 1	1,984; 1,147; 562
BR_LC	564; 563; 558	27; 27; 27	4; 4; 4	533; 532; 527

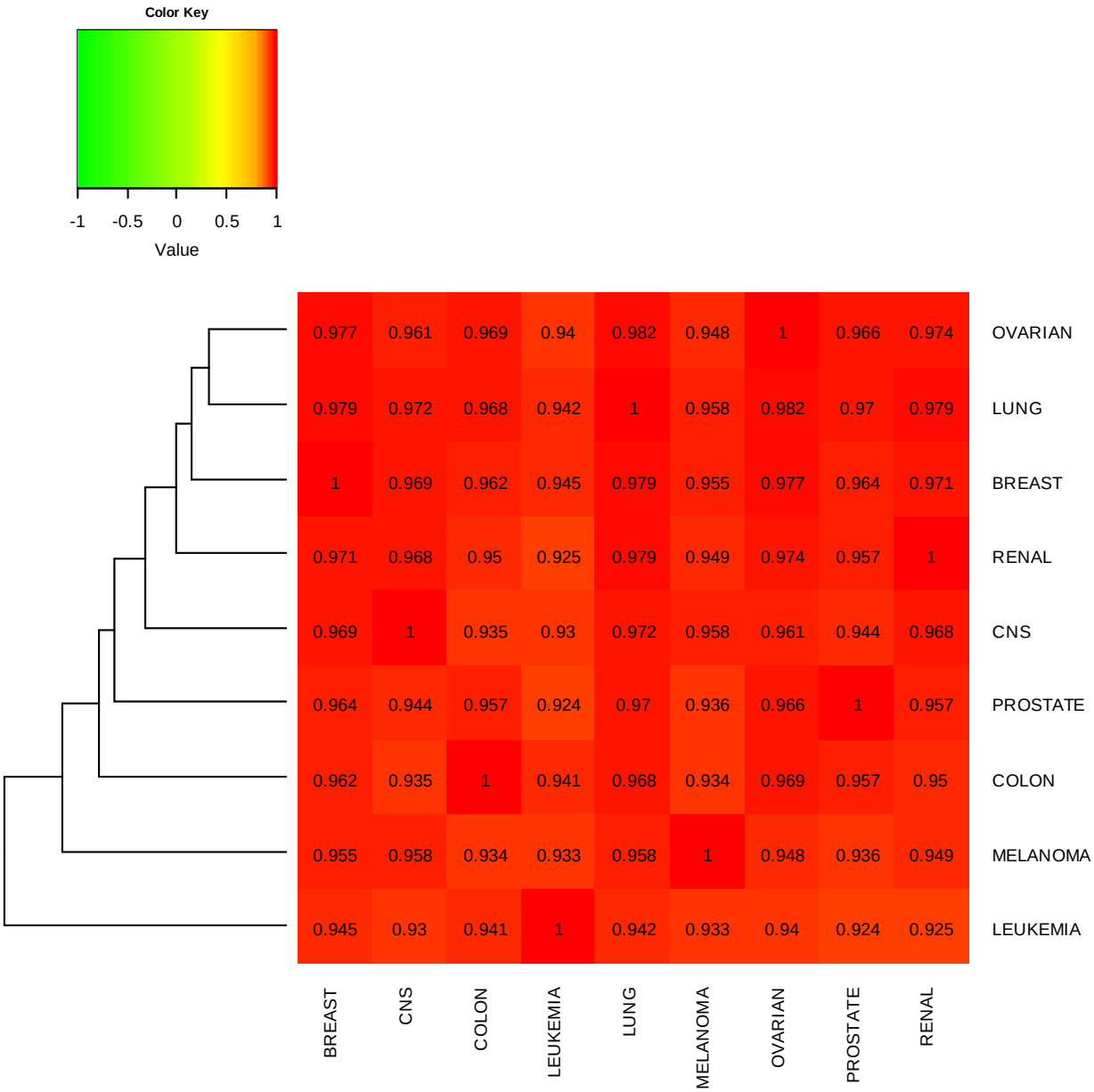
Supplementary Table 8: DEA of the protein expression data of the third data set from the proteome data with absolute log2-fold change 1, 1.4 and 2 (in the following order in the table)

Tissue comparison data	Total number of genes – log_2-fold change 1, 1.4 and 2	Number of de genes – over-expressed –log_2-fold change 1, 1.4 and 2	Number of de genes – under-expressed –log_2-fold change 1, 1.4 and 2	Number of unchanged genes –log_2-fold change 1, 1.4 and 2
BR_LE	2,236; 1,508; 830	1; 1; 1	4; 4; 4	2,231; 1,503; 825
BR_LC	2,273; 1,526; 882	1; 1; 1	2; 2; 2	2,270; 1,523; 879
BR_ME	2,568; 1,819; 1,123	14; 14; 14	8; 8; 8	2,546; 1,797; 1,101
BR_OV	2,002; 1,191; 573	0; 0; 0	0; 0; 0	2,002; 1,191; 573
BR_PR	2,156; 1,335; 713	0; 0; 0	0; 0; 0	2,156; 1,335; 713
BR_RE	2,037; 1,320; 617	0; 0; 0	1; 1; 1	2,036; 1,319; 616
CNS_CO	554; 553; 546	3; 3; 3	61; 61; 61	490; 489; 482
CNS_LE	2,188; 1,372; 717	0; 0; 0	3; 2; 2	2,185; 1,370; 715
CNS_LC	2,253; 1,567; 948	43; 42; 42	38; 37; 37	2,172; 1,488; 869
CNS_ME	2,414; 1,923; 1,118	124; 122; 122	36; 35; 35	2,254; 1,766; 961
CNS_OV	2,044; 1,258; 659	1; 1; 1	0; 0; 0	2,043; 1,257; 658
CNS_PR	2,044; 1,330; 687	11; 11; 11	6; 6; 6	2,027; 1,313; 670
CNS_RE	1,863; 1,368; 590	2; 2; 2	2; 2; 2	1,859; 1,364; 586
CO_LE	835; 833; 825	16; 16; 16	60; 60; 60	759; 757; 749
CO_LC	1,978; 1,238; 630	6; 6; 6	9; 9; 9	1,963; 1,223; 615
CO_ME	2,432; 1,670; 1,020	61; 59; 59	19; 17; 17	2,352; 1,594; 944
CO_OV	1,976; 1,281; 618	9; 9; 9	3; 3; 3	1,964; 1,269; 606
CO_PR	2,081; 1,418; 769	53; 52; 52	25; 24; 24	2,003; 1,342; 693
CO_RE	2,110; 1,344; 720	0; 0; 0	1; 1; 1	2,109; 1,343; 719
LC_LE	588; 588; 584	13; 13; 13	31; 31; 31	544; 544; 540
LC_ME	2,200; 1,482; 803	16; 16; 16	23; 23; 23	2,161; 1,443; 764
LC_OV	2,709; 1,924; 1,178	62; 59; 59	25; 25; 25	2,622; 1,840; 1,094
LC_PR	1,788; 1,106; 509	2; 2; 2	0; 0; 0	1,786; 1,104; 507
LC_RE	1,838; 1,060; 499	0; 0; 0	0; 0; 0	1,838; 1,060; 499
LE_ME	491; 489; 484	2; 2; 2	25; 25; 25	464; 462; 457
LE_OV	1,947; 1,161; 546	1; 1; 1	1; 1; 1	1,945; 1,159; 544
LE_PR	2,612; 1,889; 1,110	32; 32; 32	89; 88; 88	2,491; 1,769; 990
LE_RE	2,293; 1,754; 944	28; 27; 27	54; 53, 53	2,211; 1,674; 864
ME_OV	2,745; 2,459; 1,691	9; 9; 9	118; 118; 117	2,618; 2,332; 1,565
ME_PR	2,453; 1,715; 1,038	31; 30; 30	71; 70; 67	2,351; 1,615; 941
ME_RE	1,971; 1,229; 613	0; 0; 0	0; 0; 0	1,971; 1,229; 613
OV_PR	645; 645; 640	3; 3; 3	33; 33; 33	609; 609; 604
OV_RE	2,079; 1,338; 693	17; 17; 17	17; 17; 17	2,045; 1,304; 659
PR_RE	460; 458; 453	5; 5; 5	25; 25; 25	430; 428; 423
BR_LE	1,869; 1,075; 525	0; 0; 0	1; 1;1	1,868; 1,074; 524
BR_LC	532; 532; 527	26; 26; 26	4; 4; 4	502; 502; 497

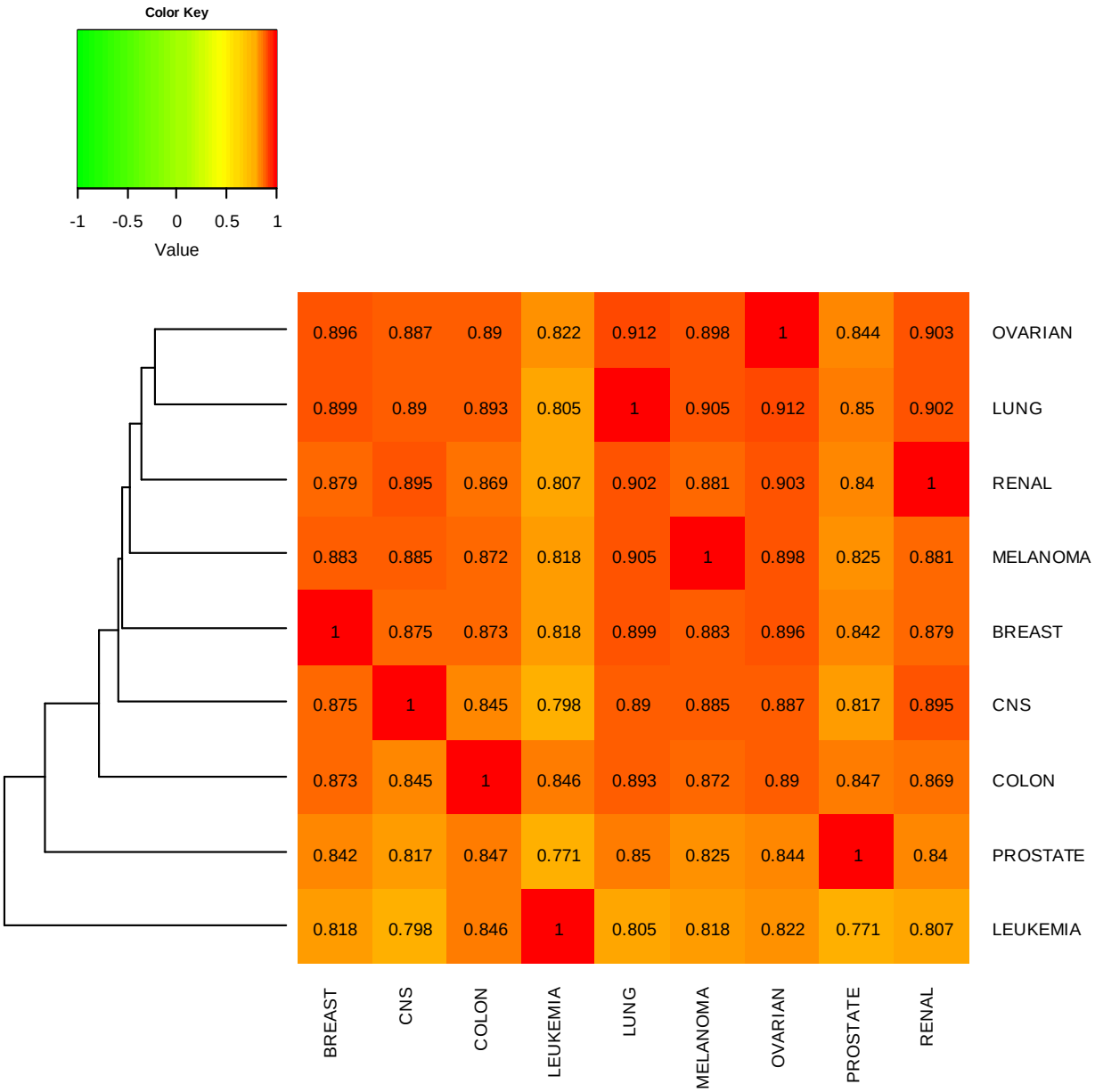
6.2 Clustering and Heatmap



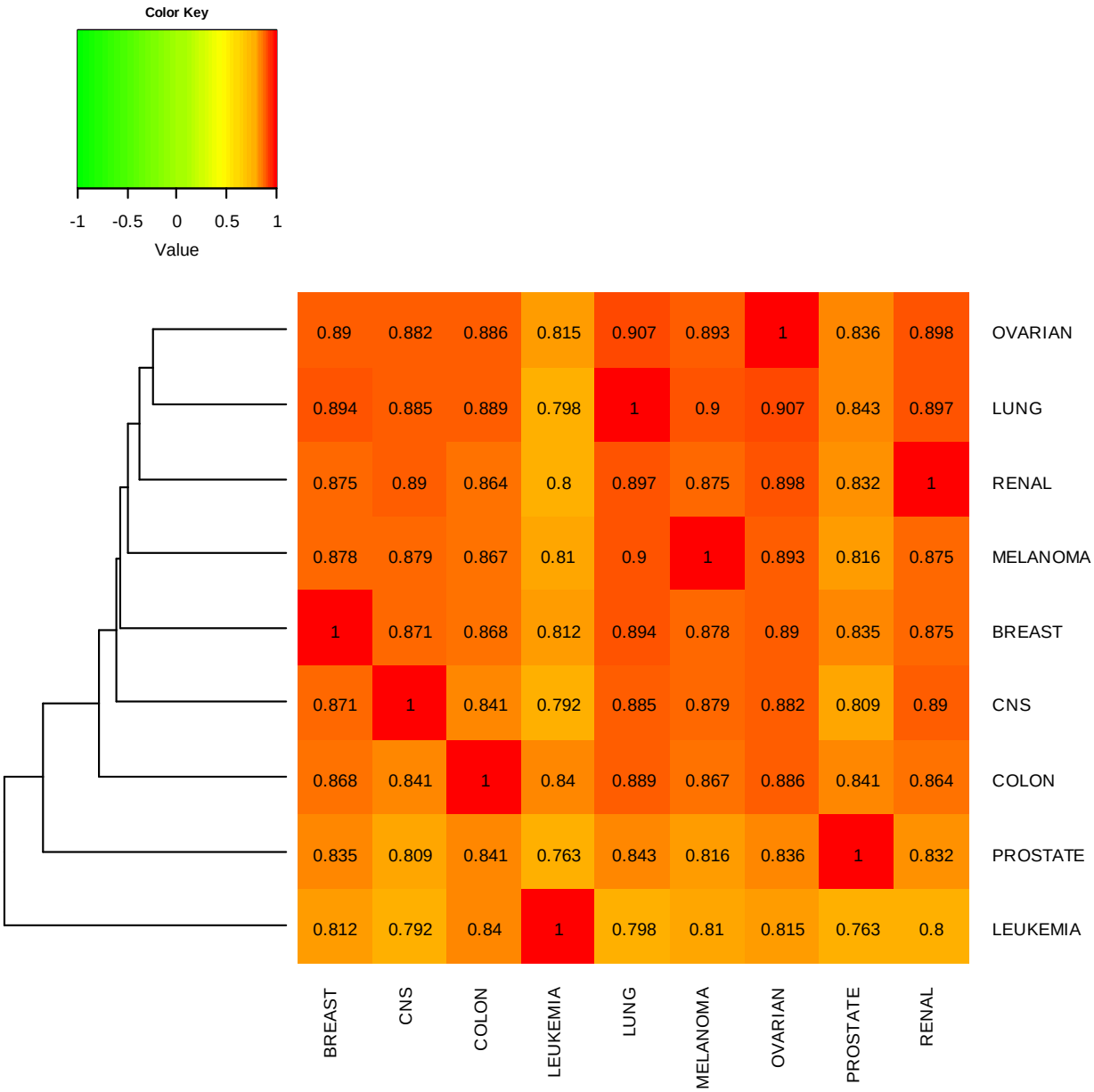
Supplementary Figure 1: Heatmap of the mean Pearson correlation coefficients of the second data set of gene expression data of 20,199 mRNA probe sets between the nine different tissues where the rows of the heatmap are arranged by the hierarchical clustering according to the average linkage method of 9 tissue origins in the NCI-60 data.



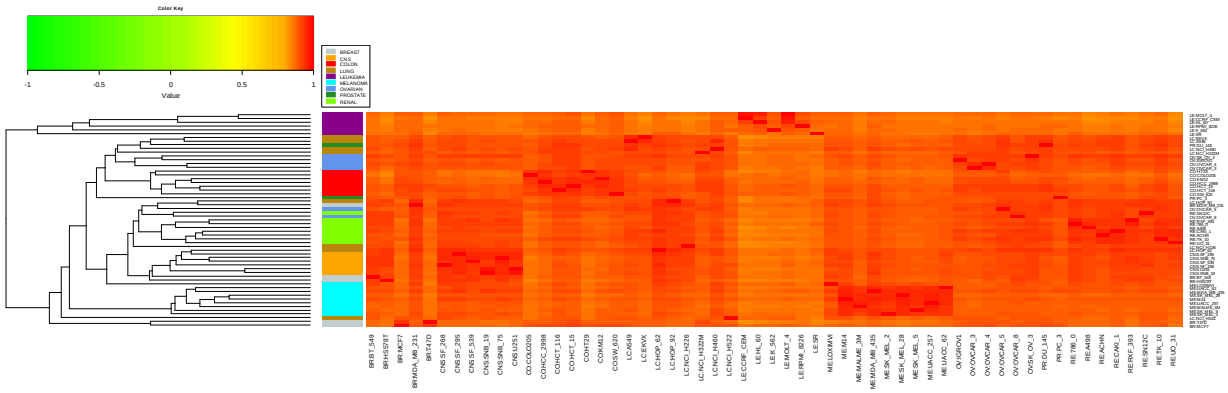
Supplementary Figure 2: Heatmap of the mean Pearson correlation coefficients of the third data set of gene expression data of 9,591 mRNA probe sets between the nine different tissues where the rows of the heatmap are arranged by the hierarchical clustering according to the average linkage method of 9 tissue origins in the NCI-60 data.



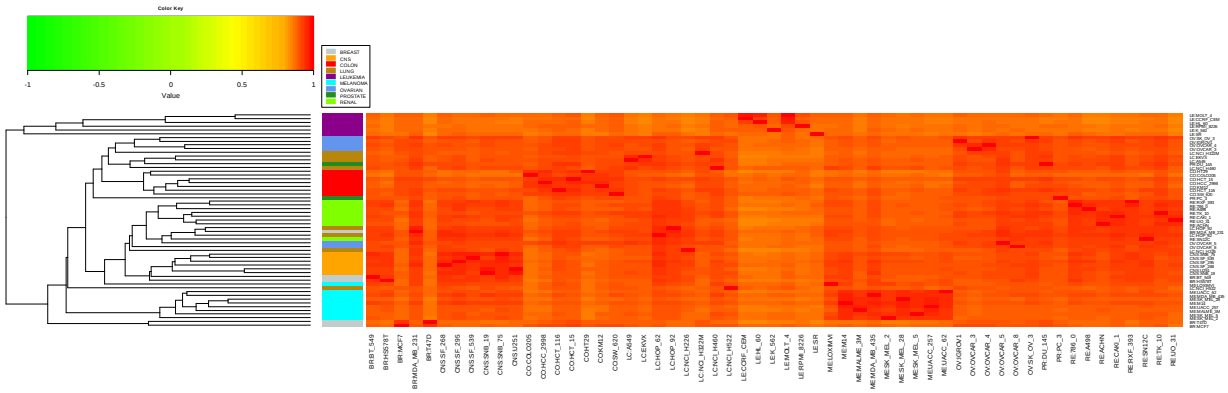
Supplementary Figure 3: Heatmap of the mean Pearson correlation coefficients of the second data set of protein expression data of 7,087 proteins where between the nine different tissues the rows of the heatmap are arranged by the hierarchical clustering according to the average linkage method of 9 tissue origins in the NCI-60 data.



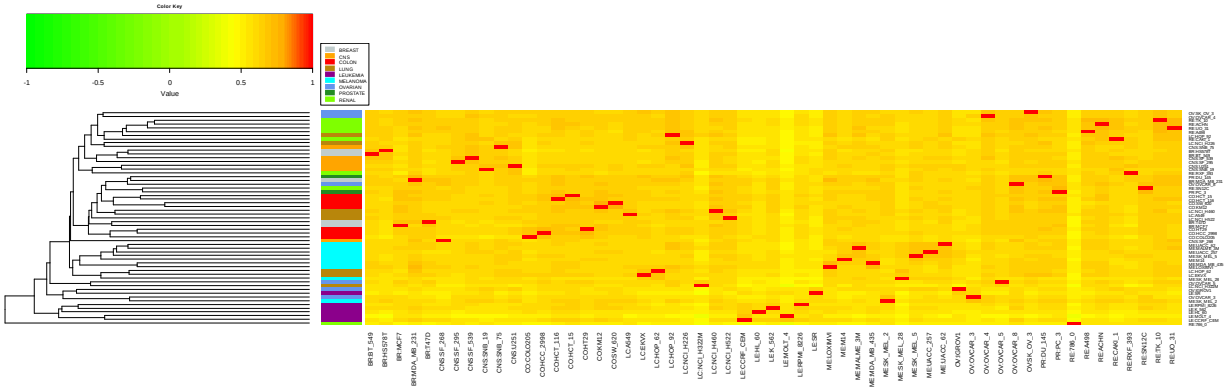
Supplementary Figure 4: Heatmap of the mean Pearson correlation coefficients of the third data set of protein expression data of 6,544 proteins where between the nine different tissues the rows of the heatmap are arranged by the hierarchical clustering according to the average linkage method of 9 tissue origins in the NCI-60 data.



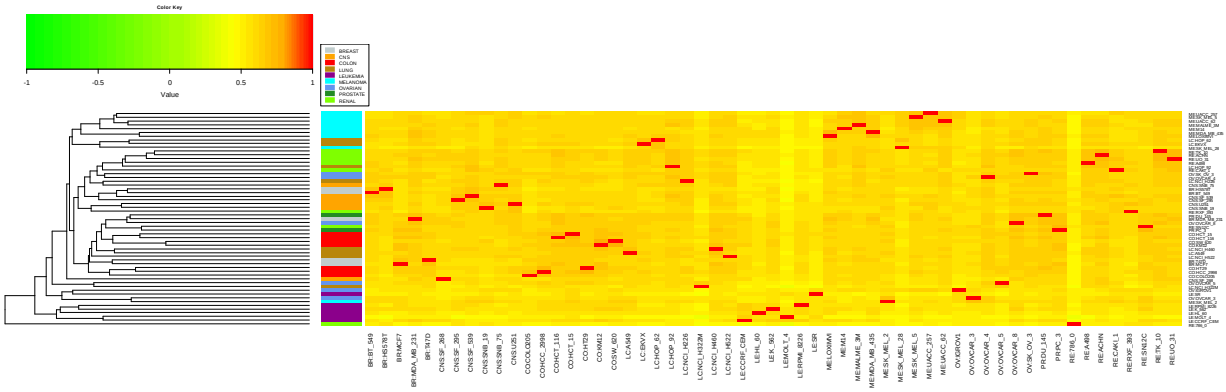
Supplementary Figure 5: Heatmap of the mean Pearson correlation coefficients of the second data set of gene expression data of 20,199 mRNA probe sets between the 57 different cancer cell lines where the rows of the heatmap are arranged by the hierarchical clustering according to the average linkage method of the 57 cancer cell lines in the NCI-60 data.



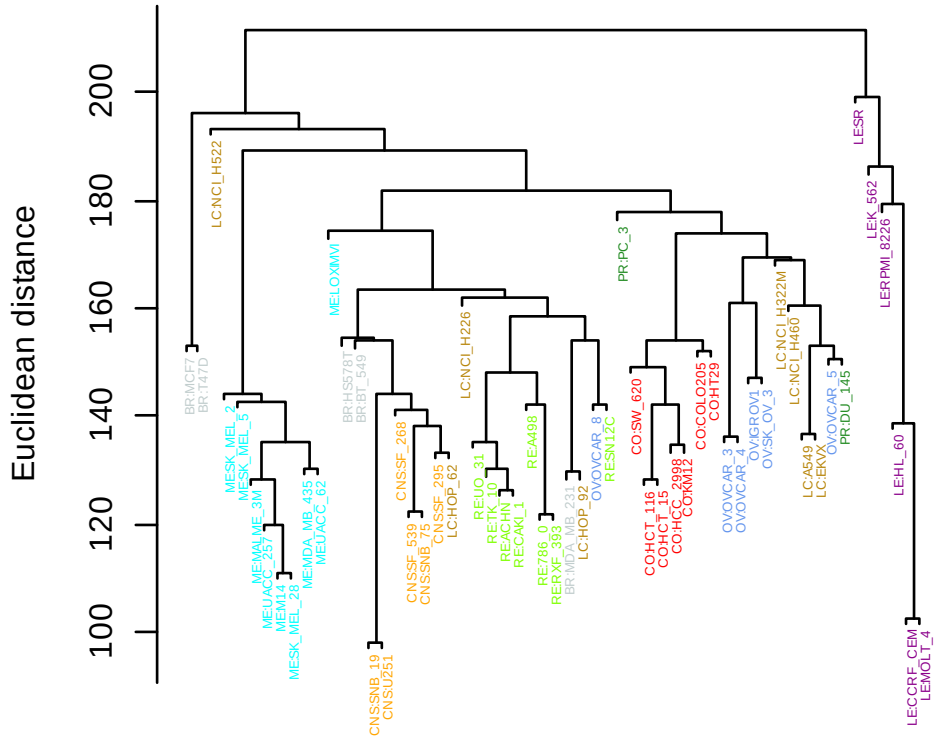
Supplementary Figure 6: Heatmap of the mean Pearson correlation coefficients of the third data set of gene expression data of 9,591 mRNA probe sets between the 57 different cancer cell lines where the rows of the heatmap are arranged by the hierarchical clustering according to the average linkage method of the 57 cancer cell lines in the NCI-60 data.



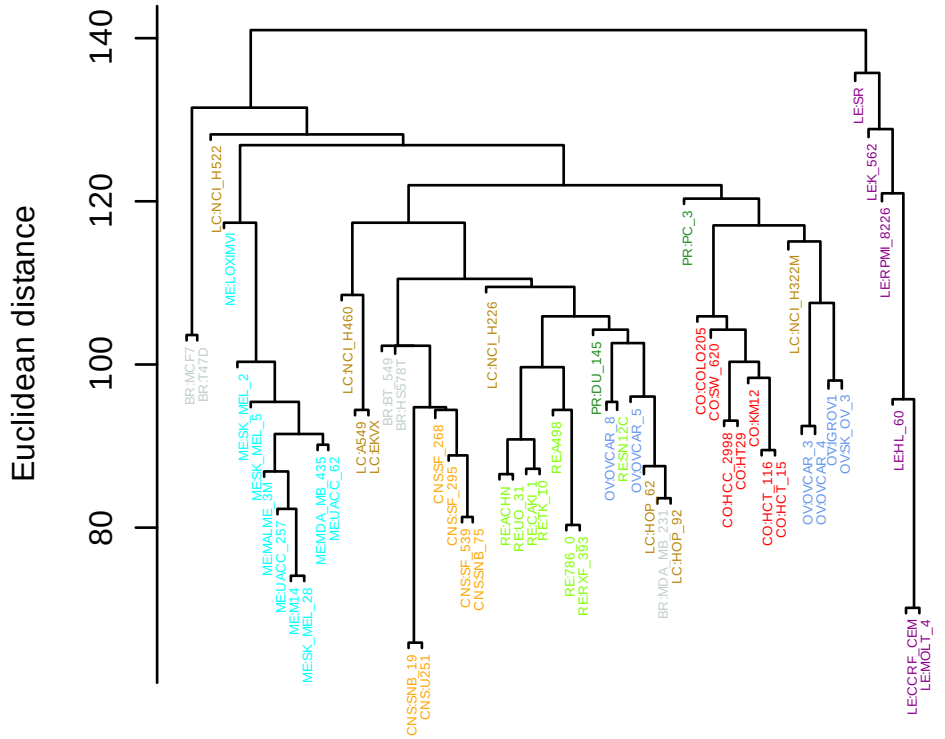
Supplementary Figure 7: Heatmap of the mean Pearson correlation coefficients of the second data set of protein expression data of 7,087 proteins between the 57 different cancer cell lines where the rows of the heatmap are arranged by the hierarchical clustering according to the average linkage method of the 57 cancer cell lines in the NCI-60 data.



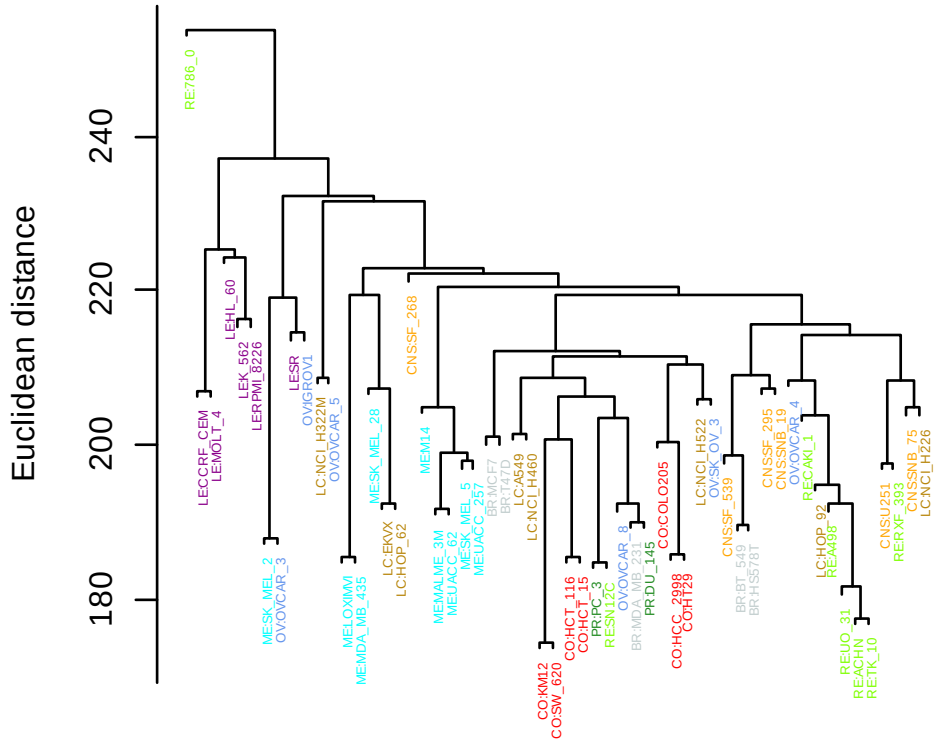
Supplementary Figure 8: Heatmap of the mean Pearson correlation coefficients of the third data set of protein expression data of 6,544 proteins between the 57 different cancer cell lines where the rows of the heatmap are arranged by the hierarchical clustering according to the average linkage method of the 57 cancer cell lines in the NCI-60 data.



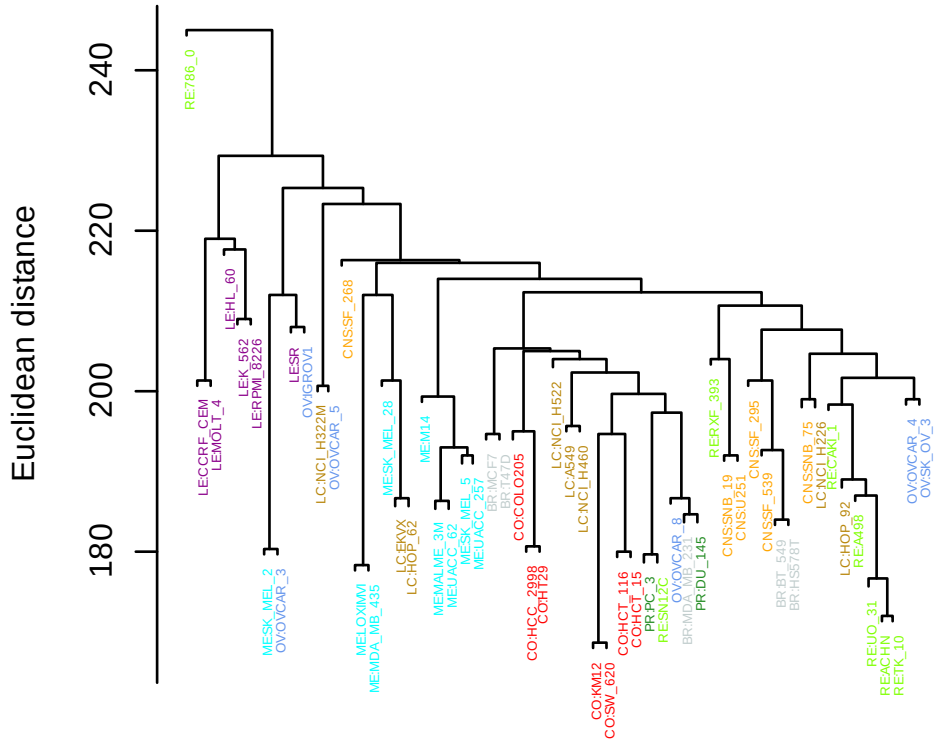
Supplementary Figure 9: Dendrogram of 57 cancer cell lines of the second data set of gene expression data of 20,199 mRNA probe sets where the samples of the dendrogram are arranged by the hierarchical clustering according to the average linkage method of the 57 cancer cell lines in the NCI-60.



Supplementary Figure 10: Dendrogram of 57 cancer cell lines of the third data set of gene expression data of 9,591 mRNA probe sets where the samples of the dendrogram are arranged by the hierarchical clustering according to the average linkage method of the 57 cancer cell lines in the NCI-60.



Supplementary Figure 11: Dendrogram 57 cancer cell lines of the second data set of protein expression data of 7,087 proteins where the samples of the dendrogram are arranged by the hierarchical clustering according to the average linkage method of the 57 cancer cell lines in the NCI-60 data.



Supplementary Figure 12: Dendrogram 57 cancer cell lines of the second data set of protein expression data of 6,544 proteins where the samples of the dendrogram are arranged by the hierarchical clustering according to the average linkage method of the 57 cancer cell lines in the NCI-60 data.

6.3 Gene set enrichment analysis

Supplementary Table 9: Results of GSEA of the second data set from HG-U133 plus 2.0 platform

Tissue comparison data	Number of DE genes	Number of members from BP GO category	Number of members from MF GO category	Number of members from CC GO category	Number of members from Reactome pathways
BR_LE	1	NA[5]	NA[1]	NA[1]	Warning[6]
BR_LC	36	127	24	22	0
BR_ME	0	0	0	0	0
BR_OV	116	589	72	93	5
BR_PR	180	494	88	73	0
BR_RE	0	0	0	0	0
CNS_CO	5	3	0	0	0
CNS_LE	7	109	16	17	0
CNS_LC	1,241	651	109	147	16
CNS_ME	37	286	50	47	0
CNS_OV	966	1,107	119	156	22
CNS_PR	689	1,044	179	159	12
CNS_RE	118	552	93	86	11
CO_LE	189	291	65	83	3
CO_LC	163	537	79	71	1
CO_ME	126	433	50	48	0
CO_OV	483	898	134	107	17
CO_PR	1,361	1,118	139	162	3
CO_RE	33	349	43	65	0
LC_LE	60	354	53	23	0
LC_ME	1,053	522	91	118	2
LC_OV	557	869	118	118	11
LC_PR	428	1,235	164	143	0
LC_RE	0	0	0	0	0
LE_ME	11	76	9	9	4
LE_OV	1	NA[1]	NA[1]	NA[1]	Warning[2]
LE_PR	1,252	936	142	135	14
LE_RE	333	1,061	119	133	20
ME_OV	165	906	73	60	13
ME_PR	1,401	882	124	186	57
ME_RE	524	898	109	142	0
OV_PR	303	451	101	77	0
OV_RE	1,278	1,149	126	165	11
PR_RE	10	80	5	5	0
BR_LE	6	217	19	2	1
BR_LC	69	425	65	38	0

[5] Error when using only 1 de gene in GSEA for GO caregories – Error: "Error in getGoToEntrezMap_db(p): genes being tested do not have corresponding GO terms"

[6] Warning in GSEA for Reactome pathways – Warning: "No gene can be mapped …. → return NULL…"

Supplementary Table 10: Results of GSEA of the third data set from HG-U133 plus 2.0 platform

Tissue comparison data	Number of DE genes	Number of members from BP GO category	Number of members from MF GO category	Number of members from CC GO category	Number of members from Reactome pathways
BR_LE	1	NA[7]	NA[3]	NA[3]	Warning[8]
BR_LC	15	60	23	16	1
BR_ME	0	0	0	0	0
BR_OV	37	228	34	35	23
BR_PR	67	253	62	24	0
BR_RE	0	0	0	0	0
CNS_CO	3	5	0	0	0
CNS_LE	2	107	2	2	4
CNS_LC	502	396	82	67	5
CNS_ME	9	116	27	14	6
CNS_OV	306	483	70	50	0
CNS_PR	254	446	110	67	0
CNS_RE	44	273	52	58	0
CO_LE	88	297	65	71	0
CO_LC	57	279	37	52	0
CO_ME	40	526	43	19	0
CO_OV	174	522	63	56	3
CO_PR	530	515	102	68	0
CO_RE	8	225	17	51	5
LC_LE	28	427	37	6	7
LC_ME	388	375	110	78	0
LC_OV	184	522	69	47	0
LC_PR	170	643	88	65	0
LC_RE	0	0	0	0	0
LE_ME	4	5	6	8	0
LE_OV	0	0	0	0	0
LE_PR	429	493	105	77	0
LE_RE	102	516	58	81	0
ME_OV	72	327	61	27	0
ME_PR	468	396	86	88	44
ME_RE	208	384	83	72	0
OV_PR	128	310	86	43	3
OV_RE	471	604	88	73	8
PR_RE	6	3	4	0	0
BR_LE	2	3	5	0	0
BR_LC	34	192	50	35	0

[3] Error when using only 1 de gene in GSEA for GO categories – Error: "Error in getGoToEntrezMap_db(p): genes being tested do not have corresponding GO terms"

[4] Warning in GSEA for Reactome pathways – Warning: "No gene can be mapped …. → return NULL…"

Supplementary Table 11: Results of GSEA of the second data set from proteome data

Tissue comparison data	Number of DE genes	Removed proteins which do not have corresponding Affymetrix ID or ENTREZ ID	Number of members from BP GO category	Number of members from MF GO category	Number of members from CC GO category	Number of members from Reactome pathways
BR_CNS	7	0	60	16	7	3
BR_CO	3	1	59	13	17	0
BR_LE	0	0	0	0	0	0
BR_LC	23	3	424	49	73	0
BR_ME	0	0	0	0	0	0
BR_OV	1	0	NA[9]	NA[7]	NA[7]	Warning[10]
BR_PR	66	4	230	79	61	0
BR_RE	3	0	84	7	17	6
CNS_CO	91	6	888	103	123	11
CNS_LE	1	0	NA[7]	NA[7]	NA[7]	Warning[8]
CNS_LC	177	13	992	149	200	11
CNS_ME	19	1	262	38	46	0
CNS_OV	4	0	211	22	6	3
CNS_PR	80	3	241	61	50	0
CNS_RE	15	0	156	27	40	3
CO_LE	13	2	260	37	28	0
CO_LC	86	2	659	97	119	0
CO_ME	82	3	521	91	70	1
CO_OV	1	1	0	0	0	0
CO_PR	44	3	281	52	50	0
CO_RE	42	3	347	62	63	1
LC_LE	94	6	455	111	140	6
LC_ME	2	0	80	8	27	20
LC_OV	0	0	0	0	0	0
LC_PR	28	0	264	44	40	6
LC_RE	2	1	NA[7]	NA[7]	NA[7]	Warning[8]
LE_ME	131	8	528	109	124	0
LE_OV	93	2	599	87	109	6
LE_PR	135	7	312	77	94	1
LE_RE	119	9	1,039	109	187	20
ME_OV	3	0	142	26	18	5
ME_PR	38	1	268	43	36	0
ME_RE	36	0	421	76	55	0
OV_PR	30	2	148	35	34	0
OV_RE	1	0	NA[7]	NA[7]	NA[7]	Warning[8]
PR_RE	31	0	167	41	51	0

[9] Error when using only 1 de gene in GSEA for GO categories – Error: "Error in getGoToEntrezMap_db(p): genes being tested do not have corresponding GO terms"

[10] Warning in GSEA for Reactome pathways – Warning: "No gene can be mapped …. → return NULL…"

Supplementary Table 12: Results GSEA of the third data set from proteome data

Tissue comparison data	Number of DE genes	Removed proteins which do not have corresponding Affymetrix ID or ENTREZ ID	Number of members from BP GO category	Number of members from MF GO category	Number of members from CC GO category	Number of members from Reactome pathways
BR_CNS	5	0	32	8	11	0
BR_CO	3	1	59	13	17	0
BR_LE	0	0	0	0	0	0
BR_LC	22	3	437	49	72	0
BR_ME	0	0	0	0	0	0
BR_OV	1	0	NA[11]	NA[9]	NA[9]	Warning[12]
BR_PR	64	4	214	76	58	0
BR_RE	3	0	84	7	17	6
CNS_CO	81	6	692	84	122	4
CNS_LE	1	0	NA[9]	NA[9]	NA[9]	Warning[10]
CNS_LC	160	12	1,075	151	183	13
CNS_ME	17	0	263	38	40	0
CNS_OV	4	0	211	22	6	3
CNS_PR	76	3	234	63	51	0
CNS_RE	15	0	156	27	40	3
CO_LE	12	2	194	27	23	0
CO_LC	80	2	625	94	104	1
CO_ME	78	3	522	81	70	5
CO_OV	1	1	0	0	0	0
CO_PR	44	3	281	52	50	0
CO_RE	39	3	337	59	52	1
LC_LE	87	6	430	97	120	8
LC_ME	2	0	80	8	27	20
LC_OV	0	0	0	0	0	0
LC_PR	27	0	179	33	37	1
LC_RE	2	1	NA[9]	NA[9]	NA[9]	Warning[10]
LE_ME	121	8	559	114	116	0
LE_OV	82	2	591	90	99	7
LE_PR	127	7	309	75	94	1
LE_RE	102	8	1,050	94	174	23
ME_OV	0	0	0	0	0	0
ME_PR	36	1	179	30	28	0
ME_RE	34	0	427	70	52	8
OV_PR	30	2	148	35	34	0
OV_RE	1	0	NA[9]	NA[9]	NA[9]	Warning[10]
PR_RE	30	0	139	29	50	0

[11] Error when using only 1 de gene in GSEA for GO categories – Error: "Error in getGoToEntrezMap_db(p): genes being tested do not have corresponding GO terms"

[12] Warning in GSEA for Reactome pathways – Warning: "No gene can be mapped …. → return NULL…"

7 List of figures

8 List of table