From DEPARTMENT ONCOLOGY AND PATHOLOGY

KAROLINSKA BIOMICS CENTER

Karolinska Institutet, Stockholm, Sweden

CANCER PROTEOMICS: METHOD DEVELOPMENT FOR MASS SPECTROMETRY BASED ANALYSIS OF CLINICAL MATERIALS

Maria Pernemalm

Stockholm 2009

All previously published papers were reproduced with permission from the publisher.

Cover artwork by Staffan Nyström

Published by Karolinska Institutet. Printed by [larserics digital print AB]

ISBN 978-91-7409-656-9

ABSTRACT

To improve cancer treatment, biomarkers for diagnostics and therapeutic guidance are desperately needed. Mass spectrometry (MS) based proteomics is one of the most promising methods for biomarker discovery. Clinical materials such as blood and tumor tissue provide an excellent starting material for biomarker discovery studies. However, at present, there are several analytical challenges related to biomarker discovery from clinical materials using mass spectrometry.

In this thesis several methodological aspects in mass spectrometry based biomarker discovery workflows are optimized, including sample preparation, sample prefractionation and data management.

In **paper I** an analytical workflow for SELDI-TOF MS of acute myeloid leukemia (AML) cells is presented including sample selection, experimental optimization, repeatability estimation, data preprocessing, data fusion, and feature selection. The study illustrates the benefit of combining the information from several data analysis methods when dealing with complex data from global proteomics analysis.

Papers II, III and **IV**, deals with analytical challenges when performing biomarker discovery studies using plasma as a starting material. The studies highlight the benefit of prefractionation on the analytical depth and in addition show the importance of identifying a large number of proteins to reach low abundant tissue leakage proteins.

Paper IV shows the added value of combining high abundant protein depletion and narrow range peptide isoelectric focusing for plasma biomarker discovery studies.

In **paper IV**, pleural effusion, a proximal fluid in lung cancer, is collected and prepared according to the same protocol as plasma; an approach that previously has not been described. The potential of using pleural effusion as discovery material is also shown.

Paper V describes a protocol for removal of blood contamination and enrichment of tumor cells from lung cancer tumor tissue. By removal of blood and stromal contaminants, twice as many proteins could be identified from lung cancer tissue, as compared with direct lysis of fresh frozen tissue.

In general this thesis highlights the importance of experimental design and optimization prior to performing biomarker discovery experiments from clinical materials, especially as clinical materials usually are limited both in amounts and numbers and the sample sets contains a high inherent variability.

LIST OF PUBLICATIONS

I. Forshed, J.; **Pernemalm, M**.; Tan, C. S.; Lindberg, M.; Kanter, L.; Pawitan, Y.; Lewensohn, R.; Stenke, L.; Lehtiö, J., Proteomic data analysis workflow for discovery of candidate biomarker peaks predictive of clinical outcome for patients with acute myeloid leukemia. *J Proteome Res* 2008, 7, (6), 2332-41.

II. **Pernemalm, M.**; Orre, L. M.; Lengqvist, J.; Wikström, P.; Lewensohn, R.; Lehtiö, J., Evaluation of three principally different intact protein prefractionation methods for plasma biomarker discovery. *J Proteome Res* 2008, 7, (7), 2712-22

III. **Pernemalm, M.**; Lewensohn, R.; Lehtiö, J., Affinity prefractionation for MS-based plasma proteomics. *Proteomics* 2009 Mar;9(6):1420-7.

IV. **Pernemalm M**.; De Petris L.; Eriksson H.; Brandén E.; Koyi H.; Kanter L.; Lewensohn R.; Lehtiö J., Use of narrow-range peptide IEF to improve detection of lung adenocarcinoma markers in plasma and pleural effusion. *Proteomics.* 2009 Jul;9(13):3414-24

V. De Petris, L.; **Pernemalm, M**.; Elmberger, G.; Bergman, P.; Orre, L.; Lewensohn, R.; Lehtiö, J., A novel method for sample preparation of fresh lung cancer tissue preparation for high resolution mass spectrometry-based proteomics. *Manuscript, submitted*

Additional Papers

De Petris, L.; Orre, L. M.; Kanter, L.; **Pernemalm, M.**; Koyi, H.; Lewensohn, R.; Lehtiö, J., Tumor expression of S100A6 correlates with survival of patients with stage I non-small-cell lung cancer. *Lung Cancer*. 2009 Mar;63(3):410-7. Epub 2008 Jul 11

Orre, L. M.; **Pernemalm, M.**; Lengqvist, J.; Lewensohn, R.; Lehtiö, J., Up-regulation, modification, and translocation of S100A6 induced by exposure to ionizing radiation revealed by proteomics profiling. *Mol Cell Proteomics* 2007, 6, (12), 2122-31.

Tan, C. S.; Ploner, A.; Quandt, A.; Lehtiö, J.; **Pernemalm, M.**; Lewensohn, R.; Pawitan, Y., Annotated regions of significance of SELDI-TOF-MS spectra for detecting protein biomarkers. *Proteomics* 2006, 6, (23), 6124-33.

CONTENTS

LIST OF ABBREVIATIONS

2DE	Two dimensional gel electrophoresis
AML	Acute myeloid leukemia
CR	Complete remission
CSF	Cerebrospinal fluid
CV	Coefficient of variation
DNA	Deoxyribonucleic acid
EDTA	Ethylenediaminetetraacetic acid
ELISA	Enzyme-linked immunosorbent assay
ESI	Electrospray ionization
ETS	Enriched tumorcell suspension
FDA	Food and drug administration
FF	Fresh frozen
FFE	Free flow electrophoresis
FFPE	Formalin fixed paraffin embedded
FTICR	Fourier transform ion cyclotron resonance
GO	Gene ontology
HPPP	Human plasma proteome project
HUPO	Human proteome organization
ICAT	Isotope-coded affinity tags
IEF	Isoelectric focusing
IHC	Immunohisto chemistry
IPG	Immobilized pH gradient
iTRAQ	Isobaric tag for relative and absolute quantification
LC	Liquid chromatography
m/z	Mass to charge ratio
MALDI	Matrix assisted laser desorption ionization
MARS	Multiple affinity removal system
MS	Mass spectrometry
MudPIT	Multidimensional protein identification technology
PSA	Prostate specific antigen
Q	Quadropole
RNA	Ribonucleic acid
RP	Reversed phase
SCX	Strong cation exchange
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SELDI	Surface enhanced lased desorption ionization
SILAC	Stable isotope labeling with amino acids in cell culture
SOP	Standard operating procedure
SRM	Selected reaction monitoring
TOF	Time-of-flight

1 BACKGROUND

1.1 PROTEOMICS

The sequence of the human genome was published in 2001 [1, 2] and is now believed to contain about 20 000 genes [3, 4]. Somewhat simplified, the genes are similar to the ingredients in a recipe. The combination of different genes makes up an individual's *genotype*, or the recipe itself. A *phenotype* on the other hand describes the observable features of an individual, such as morphology, size, physiology and behavior. Much like the genes make up the genotype, the proteins make up a large portion of the phenotype. The word *protein* comes from the Greek word 'prota', meaning 'of primary importance'.

In analogy to the human *genome* there is also a corresponding human *proteome*. The term 'proteome' was first introduced by Marc Wilkins in 1994 and was subsequently published in 1995 [5]. Wilkins used it to describe the entire protein complement of a genome, a cell, a tissue or an organism. As of today the entire human proteome has not been mapped, but it has been estimated that a single human cell contains on average 100 000 proteins [6].

In the so called post-genome era several different *–omics* techniques have emerged, aiming at studying entire *–omes*, rather than one molecule at the time. Depending on what types of molecules are studied, different –omics fields have been defined; proteomics (proteome/proteins), genomics (genome/genes), metabolomics (metabolome/metabolites), lipidomics (lipidome/lipids) etc.

There are several conceptual differences when studying the human proteome as compared with the human genome and they all comprise analytical challenges in proteomics.

First, there is not a one to one relationship between the number of genes and the number of proteins, as proteins come in different splice variants and in addition undergo post-translational modification, where sugars, phosphates and other molecules are added to the protein structure [7].

Second, not all proteins are present in all cells, and further, there is also a large difference in protein abundance, spanning over up to ten orders of magnitude in human plasma [8]. This is of particular importance for the analysis of proteins contra genes as there is no amplification technique available for proteins, as polymerase chain reaction (PCR) is available for amplification of gene materials.

Third, proteins are chemically more heterogeneous and diverse as a group than DNA and RNA. Proteins differ largely in solubility, stability, size and p*I*.

Taken together, these challenges often cause a biased discovery of high abundant and easily observed proteins in proteomics experiments. At present only one organism's proteome has been almost completely sequenced; yeast [9]. A human proteome detection and quantitation project is currently being discussed [10].

In proteomics, two-dimensional gel electrophoresis (2DE) together with mass spectrometry has traditionally been the most common combination of analytical techniques. Using 2DE proteins are separated in two dimensions based on p*I* and size. The gel is subsequently stained and protein spots of interest are identified using mass spectrometry. There are several good reviews that cover the 2DE technology and its use in proteomics [11-13].

Affinity based proteomics methods are also widely used, either by antibody arrays, where hundreds of antibodies can be immobilized on a slide and used as a multiplexed enzyme linked immunosorbent assays (ELISA), or reversed phase arrays were the samples (fluid, cells or cell lysates) are immobilized and the antibody is subsequently applied, or by tissue microarray, which enables parallel analysis of hundreds of formalin fixed paraffin embedded (FFPE) samples [14-19]. Affinity based methods also include the study of protein-protein-interactions, or *interactomes [6, 20]*.

Recently, mass spectrometry based workflows have become increasingly common, much due to advances in mass spectrometry technologies, and the possibility to a higher level of automation. Mass spectrometry based proteomics technologies are at present used to study protein identification, modification, quantification and localization (imaging).

1.2 MASS SPECTROMETRY

Mass spectrometry has become the number one analytical tool in many proteomics studies. In brief, a mass spectrometer separates ions in the gas phase based on their mass to charge ratio (m/z). Any mass spectrometer is essentially build up by three major parts; an **ion source**, a **mass analyzer** and a **detector** (fig 1).

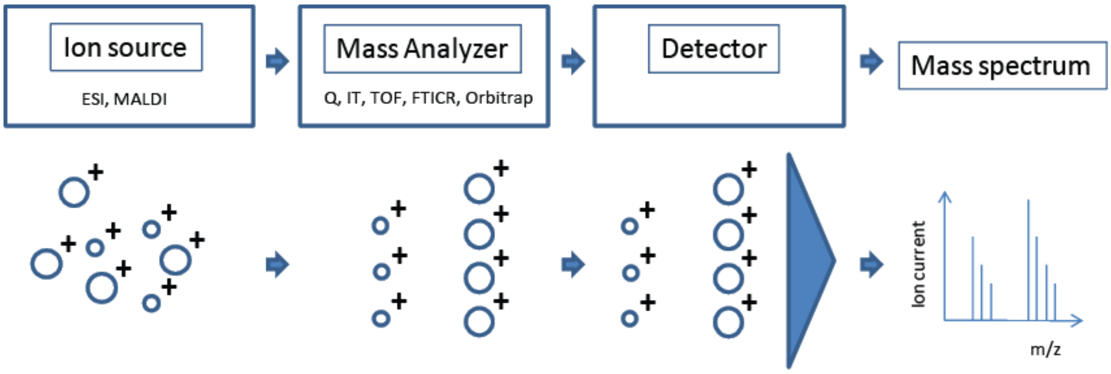


Figure 1) Schematic overview of a mass spectrometer

In the **ion source** the analytes are ionized and brought into the gas phase. In proteomics the most frequently used ion sources are either electrospray ionization (ESI) [21, 22] or matrix assisted laser desorption ionization (MALDI) [23, 24]. ESI ionizes the analytes from a solution and is therefore easily coupled on line to liquid chromatography (LC). In MALDI the sample co-crystallizes with a matrix and is subsequently pulsed with a laser, which ionizes and vaporizes the analytes.

Once in the gas phase the analytes are separated based on their m/z in the **mass analyzer**. *Quadropole* (Q), *Quadropole Ion trap* (IT), *Time of Flight* (TOF), *Fourier Transform Ion Cyclotron Resonance* (FTICR) and *Orbitrap* mass analyzers can all be used together with both MALDI and ESI ion sources and will be briefly described below.

The *quadropole* analyzer works much like a mass filter, where only a single mass/charge ratio is passed through the system at any time. The mass selectivity is created by the use of oscillating electrical fields, which stabilize or destabilize the paths of ions. To scan a wide mass range the oscillating electrical fields can be changed rapidly [25-27].

In the *quadropole ion trap* analyzers the ions are introduced to the mass analyzer in a pulsing mode, as opposed to normal quadrupoles in which ions continually enter the mass analyzer [25]. In the ion trap ions that enter the mass analyzer are detained or trapped. In essence, an ion will be stably trapped depending on the mass/charge ratio. A *linear quadrupole ion trap* (LTQ) is similar to a quadrupole ion trap, but it has an extended volume in the ion trap to increase the sensitivity.

The *time of flight* mass analyzers measures the time it takes for the ions to travel through a flight tube. The velocity of the ions is proportional to the mass, where small molecules travel faster [26, 28]. An electric field is used to accelerate the ions into the free flight zone in the flight tube.

The *fourier transform ion cyclotron resonance* analyzer measures mass by detecting the image current produced by ions cyclotroning in a magnetic field[29]. The ions which are affected by a magnetic field move at a given cyclotron frequency depending on their m/z and this is subsequently measured. By using Fourier transformation the frequency is converted to a mass to charge value.

The *Orbitrap* mass analyzer is very similar to a FTICR analyzer, but is non-magnetic and utilizes an electrostatic field instead of a magnetic field to separate the masses [30-33]. The Orbitraps that are commercially available are LTQ-Orbitraps, thereby combining the benefits of an LTQ instrument (speed, large trapping capacity, MS^n capability and versatility) with the benefits of an FTICR instrument (high mass accuracy, high resolving power, high sensitivity and high dynamic range). In addition it is more compact, less costly and easier to maintain than a LTQ-FTICR instrument [34]. The Orbitrap therefore gained much attention in the proteomics field when it was introduced in 2005 [35].

Once separated by m/z, the ions hit the **detector** and it registers the number of ions at any given m/z value. Most commonly microchannel plate detectors are used. In FTICR and Orbitrap mass spectrometers, the detector consists of a pair of metal surfaces, which the ions pass near when oscillating in the mass analyzer. Common for all detectors, the signal is converted to a mass spectrum with m/z on the x-axis and ion count/intensity on the y-axis.

There are basically two conceptually different workflows in mass spectrometry based proteomics, *top-down proteomics* and *bottom-up proteomics*. The concept of top-down and bottom-up approaches is traditionally used in software design and are basically two strategies for information processing. Simplified, in a top-down approach one starts from an overview and then go into details, and in a bottom-up approach one starts with the details and from then build up the overview.

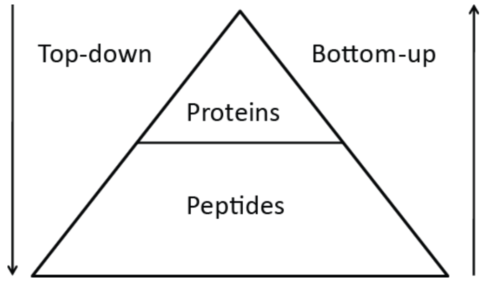


Figure 2. Conceptual overview of top-down and bottom-up strategies in proteomics.

1.3 TOP-DOWN PROTEOMICS

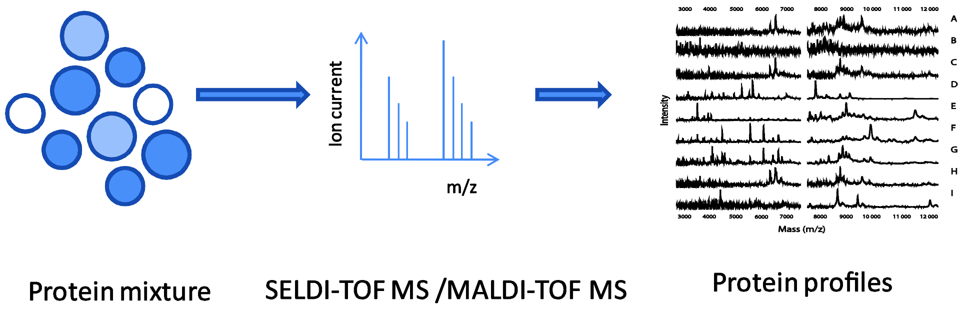


Figure 3. Schematic overview of mass spectrometry based top-down proteomics

In top-down proteomics approaches intact protein samples are analyzed directly either through classical two-dimensional electrophoresis, by antibody based methods such as antibody arrays, or by mass spectrometry. In mass spectrometry based top-down proteomics MALDI-TOF MS or SELDI-TOF MS (surface enhanced laser desorption ionization) are the most widely used technical platforms.

In top-down SELDI-TOF or MALDI-TOF analyses, the mass spectrum gives no information about the identity of the proteins, but only the relative abundance of different masses. The protein abundance pattern or *protein profile* is then analyzed and selected masses of interest can be purified and identified. There are also mass spectrometry based top-down approaches were intact proteins are subjected fragmentation and the identities of the individual proteins are obtained [36, 37], however as these techniques are rarely applied to large-scale proteome wide analysis, they will not be discussed here.

In MALDI-TOF profiling the sample can be directly applied to a MALDI target and the protein or peptide pattern can be used to distinguish between different biological states [38]. Another MALDI based top-down approach is MALDI imaging, where tissue slides are covered with matrix and analyzed directly in the mass spectrometer. This approach gives a unique spatial information about masses, which can be either protein or peptides or drug molecules [39-42].

SELDI-TOF MS is a chip based MALDI technique where the sample is analyzed directly on a selective solid-phase affinity surface [43]. Chromatographic surfaces like hydrophobic/reversed phase, anionic exchange, cationic exchange, hydrophilic/normal phase, or metal ion affinity are most commonly used, but more specific biological molecules can also be coupled to the surface. **Paper I** in this thesis is an example of a SELDI-TOF top-down proteomics study, where protein lysates from cells from patients diagnosed with acute myeloid leukemia are analyzed with SELDI-TOF-MS to detect prognostic markers.

1.4 BOTTOM-UP PROTEOMICS

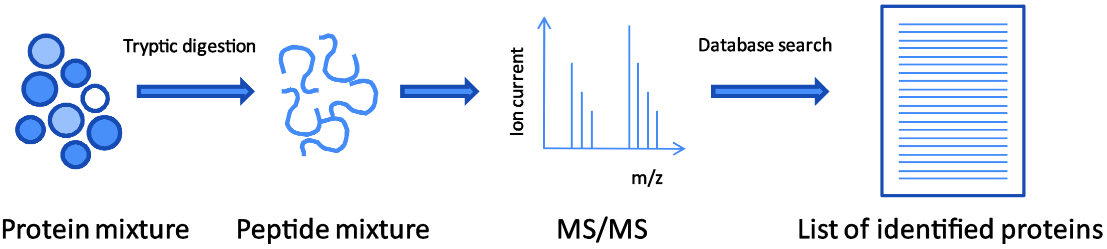


Figure 4. Schematic overview of bottom-up proteomics

The bottom-up approach has become the far most common workflow in mass spectrometry based proteomics during the last few years. Also known as shotgun proteomics (in analogy to shotgun sequencing in genomics), this approach is based on enzymatic cleavage of proteins into peptides, most usually by trypsin. The enzymatic cleavage is performed to facilitate ionization and fragmentation and subsequent identification of the proteins.

To reduce the complexity of the peptide mixture the peptides are subjected to chromatographic separation prior to mass spectrometry analysis. Reversed phase separations dominate the setups, as the reversed phase mobile phase is compatible with ESI and MALDI ionization, thereby enabling direct coupling up front to the mass spectrometer. The peptides are then analyzed by tandem mass spectrometry, where the peptide sequences are determined. In brief the peptides are separated according to mass, partially fragmented into amino acids and the fragment spectra together with the precursor mass is then used to determine the amino acid sequence of each peptide. The identified peptides are then searched against protein sequence data bases to match the peptide sequences with known protein sequences. A selection of commonly used tandem mass spectrometry set-ups in proteomics are reviewed in [44].

The bottom-up proteomics approach is commonly used together with a wide range of samples, up front separation techniques, and downstream data analysis tools. In this thesis, several different bottom-up approaches are used (**paper II, IV** and **V**) where the common denominator is that proteins are digested with trypsin and then separated off-line, using reversed phase chromatography, before identification of the peptides using MALDI-TOF/TOF mass spectrometry.

1.5 PREFRACTIONATION

Technical differences between individual mass spectrometers related to sensitivity and mass accuracy greatly influence the performance of proteomics analyses.

In addition, the level of sample complexity influences the performance of the mass spectrometry analysis. High sample complexity in proteomics samples is characterized by large number of chemically diverse analytes and a high dynamic range of concentrations. These sample characteristics are of analytical importance as they are influenced by technical limitations in mass spectrometry.

For example, dependent on chemical characteristics, all analytes do not have the same ionization properties and therefore, in a complex sample, it is difficult to obtain optimal ionization for all analytes. In addition, the ionization process is competitive, which is important especially when analyzing a large number of analytes with high dynamic range of concentrations. In tandem mass spectrometry, the fragmentation efficiency is also different between various analytes. Last, mass spectrometers have limited dynamic range of detection (usually between three to four orders of magnitude), thereby limiting the sensitivity and the quantification of the analysis of complex samples.

To overcome these analytical challenges the most common approach is to reduce the sample complexity by prefractionation. As most mass spectrometers are coupled to a liquid chromatography system, either online (directly coupled) in ESI mass spectrometry (LC-MS) or offline in MALDI (LC-MALDI) there is already one inherent chromatographic fractionation step of the sample, hence the term *prefractionation*; prior to LC-MS. Prefractionation can be performed either on a protein level or on a peptide level or using a combination of the two, and a selection of common methods will be briefly presented below.

1.5.1 Protein level

Protein pre-fraction can be performed both prior to top-down and bottom-up proteomics analyses.

Classical liquid chromatography methods such as reversed phase, ion-exchange as well as size exclusion have all been used to separate proteins based on their physio-chemical properties prior to mass spectrometry [45-47].

Affinity based prefractionation aims at enriching specific sub-groups of proteins of interest such as glycosylated proteins [48, 49] or specific interaction partners [50, 51] or to remove less interesting proteins, using for example antibody based high abundant protein depletion to remove high abundant proteins from plasma [52, 53].

Separating proteins by their p*I*, as conducted in the first dimension in 2DE, can also be performed prior to mass spectrometry analysis, but then preferably in solution using for example the OFFGEL system [54], free-flow electrophoresis (FFE) [55] or the rotofor [56].

Similarly, the second dimension in 2DE, SDS-PAGE, has also been use as a prefractionation strategy, separating proteins based on their size [57].

1.5.2 Peptide level

In bottom-up approaches the sample complexity is increased by enzymatic cleavage, therefore prefractionation on the peptide level is particularly valuable.

One of the most frequently used set-ups in shot-gun proteomics is a two-dimensional orthogonal peptide separation combining strong anion exchange (SCX) and reversed phase (RP). Denoted MudPIT (multidimensional protein identification technology) this method was first described by Yates *et al*. [58, 59].

As in prefractionation on the protein level, affinity enrichment can also be applied on the peptide level to enrich for sub-groups of interest. This can be performed to enrich for peptides containing post-translational modifications, for example using metal ion affinity [60-63] and antibodies [64-66] to enrich for phosphorylated peptides, or using lectin affinity [67] and hydrazide chemistry [49] to enrich for glycosylated peptides. Recently a novel peptide affinity method was described using group-specific anti-peptide antibodies. The Triple-X proteomics antibodies can be designed to enrich for various classes of peptides with identical terminus [68, 69].

Isoelectric focusing on the peptide level has been applied to proteomics using both gel-based sytems [70-75] and in-solution systems such as FFE[55] and OFFGEL[76]. Narrow range peptide isoelectric focusing is one of the core techniques used in this thesis and is discussed in more detail under the materials and methods section.

1.6 QUANTIFICATION AND DATA ANALYSIS

To be able to measure quantitative differences in protein abundance by mass spectrometry several quantification methods have been developed. In global protein analysis these quantification methods are, in general, relative - comparing the individual proteins or peptides between the experiments, rather than giving an exact concentration of the protein. There are, however, targeted mass spectrometry methods for absolute quantification of proteins, such as selected reaction monitoring (SRM), which is discussed in more detail in section 1.7.1.

In global protein analysis there are two principally different approaches for quantification; label free quantification and quantification using isotopic labels.

Quantitative global mass spectrometry analyses generate extremely large datasets, making manual interpretation of the data nearly impossible. Instead, most data analysis steps, from peak detection in individual mass spectrum, to identification, quantification, and statistical and biological interpretation of the data involve computational data analysis tools. Computational proteomics is an area within proteomics which blends mathematical, computational and statistical algorithms to address key issues related to protein identification and quantification from raw mass spectrometry data. This is a large field within proteomics which is not in the scope of this thesis, however, some basic data analysis concepts of importance for this thesis will be introduced below. For recent reviews on bioinformatics, computational proteomics and data analysis in mass spectrometry based proteomics please see [77-79].

1.6.1 Quantification

In top-down proteomics approaches such as MALDI-TOF MS and SELDI-TOF MS the quantification is usually label-free and based on direct comparison of peak intensities (peak height) across spectra.

In bottom-up approaches label free quantification is slightly different as several peptides per protein are identified and subsequently quantified. In addition, the LC step usually involves individual peptides eluting over several mass scans/spectra. To be able to capture as much of each m/z intensity signal as possible, the individual mass spectrometric peak areas are usually integrated over the chromatographic time scale and compared between samples, often by creating three dimensional maps with the chromatographic time scale on the x-axis, the ion intensity on the y-axis and the m/z values on the z-axis.

Another label-free quantification method for LC-MS/MS data is spectral counting, where the number of times that peptides from a certain proteins are fragmented is used as a proxy for the proteins abundance [80, 81].

Quantification using isotopic labeling can be divided into the following subgroups; *metabolic labeling*, *enzymatic labeling* and *chemical modification labeling*. One advantage with stable isotope labeling is that it enables pooling of samples, so that the quantitative analysis is performed within one spectrum and not across spectra. In addition, technical variability between samples is avoided by pooling and the number of samples to be analyzed with mass spectrometry is reduced.

The most common *metabolic labeling* strategy is SILAC; stable isotope labeling by amino acids in cell culture [82, 83]. In the SILAC workflow the cell medium contains either non-labeled or isotopically labeled ‘heavy’ amino acids. Basically all amino acids could be labeled, but the use of an essential amino acid, which does not metabolize to a different amino acid, is most desirable in order to avoid a mixture of labeled amino acid products. Cell medium containing normal amino acid is used as control, and then the samples can be grown in medium containing for example $^{15}N_2$-lysine (+2 Da), $^{15}N_4$-arginine (+4 Da), $^{13}C_6$-$^{15}N_2$-lysine (+8 Da) and $^{13}C_6$-$^{15}N_4$-arginine (+10 Da). Arginine and Lysine are isotopically labeled to make sure that all tryptic peptides contain at least one labeled amino acid. The relative quantification is then performed by comparing the intensity of the labeled and non-labeled peptides in the MS spectrum.

The first *chemical labeling* technique for mass spectrometry based proteomics was described in 1999 and denoted ICAT; isotope-coded affinity tag [84]. The ICAT tag is covalently coupled to the cystein residues in the peptides. The ICAT tag contains either zero, or eight, deuterium atoms as well as a biotin tag for the purification of the labeled peptides. Cysteins are relatively rare in proteins, and therefore enriching for the labeled peptides also comprise a reduction of the complexity in the samples. As in SILAC, ICAT quantification is performed on the peptide level.

ITRAQ labeling (isobaric tags for relative and absolute quantification) is conceptually different from SILAC and ICAT, as fragmented *reporter ions* from the tag are used for quantification in MS/MS mode [85, 86]. As stated in the name, iTRAQ labels are isobaric i.e. have the same mass, and in addition the same chromatographic properties. The iTRAQ label is covalently bound to free amines in the peptides, which means that every tryptic peptide will contain at least one label on the N-terminus of the peptide and usually more as trypsin cleaves after lysine and arginine, which both contains free amines. What distinguishes the individual tags are their fragmentation patterns in MS/MS, giving rise to reporter ions of different masses that can be quantified in the MS/MS spectrum. At present up to eight samples can be labeled and quantified in parallel using the iTRAQ labels. For a more detailed description on iTRAQ labeling see figure 6.

In addition to the labeling technologies described here there are also less frequently used isotopic labeling methods available such as isotope coded proteomics labels (ICPL) [87] and the 2-nitrobenzenesulfenyl (NBS) reagent [88].

In *enzymatic labeling* Glu-C or trypsin is used to incorporate ^{18}O during protein digestion [89, 90]. However, as it is rare that all peptides are incorporated with ^{18}O, this technique is usually not applied on large scale experiments.

1.6.2 Biological interpretation

In mass spectrometry based proteomics the result of the analysis is often a long list of identified and quantified proteins, by itself providing little insight into the biological state investigated. To assist functional analysis and contextualization of the protein catalogue several bioinformatics tools are available.

Gene ontology [91] is an annotation database, where standardized terms are grouped under three main ontologies; cellular component, biological process and molecular function. The ontology terms are assigned to individual proteins by collaboration with numerous databases such as the FlyBase (*Drosophila*), the *Saccharomyces* Genome Database (SGD) and the Mouse Genome Database (MGD). A complete list of the databases is available on www.geneontology.org. The gene ontology annotation database can easily be used to identify over and underrepresentation of terms. This is used to obtain initial insights in the sample characteristics, for example regarding sampling biases (such as underrepresentation of membrane proteins). Subgroups of proteins that are differently expressed between samples are also commonly analyzed to reveal functional clues about the system studied. Similarly, the KEGG pathway database[92] can be used to look for over and underrepresentation of specific pathways. As most proteins carry out their functions within a network of interactions, much effort has been put in to describing and characterizing protein interaction networks [93-95]. Taking advantage of this knowledge on protein interaction networks and signaling pathways, proteomics data can be used to pinpoint activation or de-activation of specific signaling cascades.

There are numerous software tools available for functional analysis of proteomics data, both commercial and academic. DAVID[96], PANTHER[97], ProteinCenter (Proxeon), Biobase (Biobase International), Ingenuity pathway analysis (Ingenuity systems), Pathway search engine (PSE) [98] and FunCoup [99] are all examples of search tools that can be used to organize proteins into groups of molecular functions, protein families, biological processes, and pathways to discover common threads underlying the proteins of interest.

1.7 BIOLOGICAL VALIDATION

Global protein analysis is labor intense and expensive and is therefore usually selectively performed on a limited number of samples. Statistically this is problematic as the number of variables by far exceeds the number of samples. This calls for thorough validation of the results, both of the protein identification results and the quantitative results. Changing technical platform, reducing the number of proteins to be monitored and increasing the number of samples is desirable at this stage. Classical molecular biology techniques such as western blot, immunohistochemistry and ELISA as well as functional analyses like siRNA and over expression are all regularly used to validate proteomics results. In addition, targeted proteomics technologies can be used as high throughput validation tools.

1.7.1 Targeted proteomics techniques

Targeted proteomics techniques can be applied as validation techniques following global proteomics analysis. These technologies can be either mass spectrometry based or antibody based.

Quantitative analysis of peptides can be performed using selected reaction monitoring (SRM) by triple quadropole mass spectrometry (Q-Q-Q). In peptide SRM selected peptides are fragmented and specific fragments are used for quantification. In addition to validating the identification, absolute quantification can be performed by SRM using stable isotope standards [10, 100-102]. Up to 50 different proteins have been successfully analyzed in parallel from plasma using peptide SRM [103].

Antibody based high-throughput methods are regularly used to validate proteomics results. Tissue microarrays [104] provide a powerful technique to analyze paraffin embedded samples in a high-throughput manner. By taking small core biopsies from the donor blocks and inserting them into a common recipient block, immunohistochemical (IHC) staining can be performed on hundreds of samples at the same time. In addition to validating the identification and the quantification the tissue samples also provide additional information on the cellular localization. Cell lysates and biological fluids can also be analyzed in a high throughput manner using reversed lysate arrays [105, 106] or antibody microarrays [107-109], where either the sample or the antibodies are printed on glass slides. Quantification is usually performed with a fluorescent labeled secondary antibody.

1.8 CANCER PROTEOMICS

The transformation of a normal cell into a cancer cell is a multi-step process, which has been described well in Hanahan and Weinbergs review from 2000 “Hallmarks of cancer” [110]. Hanahan and Weinberg describe six types of genetic alterations essential for development of malignant cancer cells; limitless replicative potential, sustained angiogenesis, evasion of apoptosis, self-sufficiency in growth signals, insensitivity to anti-growth signals and tissue invasion and metastasis. In most cases, cancer development is a slow process and is governed under Darwinian rules of selection, where cells with the capability to proliferate are continuously selected for [111]. Fewer than 10% of all cancers are caused by Mendelian inheritance.

There are basically two different starting points when studying cancer using proteomics methods; one, to gain novel insights into cancer biology and two, to try to identify clinically useful biomarkers. Somewhat simplified, studies dealing with cancer biology usually are performed in model systems, such as cell lines or animal models and biomarker discovery studies often explore clinical materials such as blood or tumor tissue.

In this thesis two different malignancies have been studied; acute myeloid leukemia (AML) and lung cancer.

AML is the most common type of leukemia and is characterized by uncontrolled growth of cells from the myeloid linage in the bone marrow. Approximately 300 new cases of AML are diagnosed in Sweden per year, and most of the patients are around 60 years old. The patients are treated with chemotherapy and normally respond well to initial treatment and go into a period of complete remission (CR). However, most patients relaps and develop resistance to treatment. The five year survival in AML is approximately 20%.

Lung cancer is the fourth most common cancer in Sweden and the most common cause of cancer related death. Approximately 3000 new cases are diagnosed every year. Lung cancer is divided into two subtypes; small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). NSCLC is the most common subtype and is further divided into three histologies; squamous cell carcinoma, adenocarcinoma and large cell carcinoma. The only curative treatment for lung cancer at present is surgery, and that can only be performed at an early stage of the disease. Additional treatments include radiotherapy and chemotherapy, but the 5-year survival remains low, below 15%.

1.9 BIOMARKERS

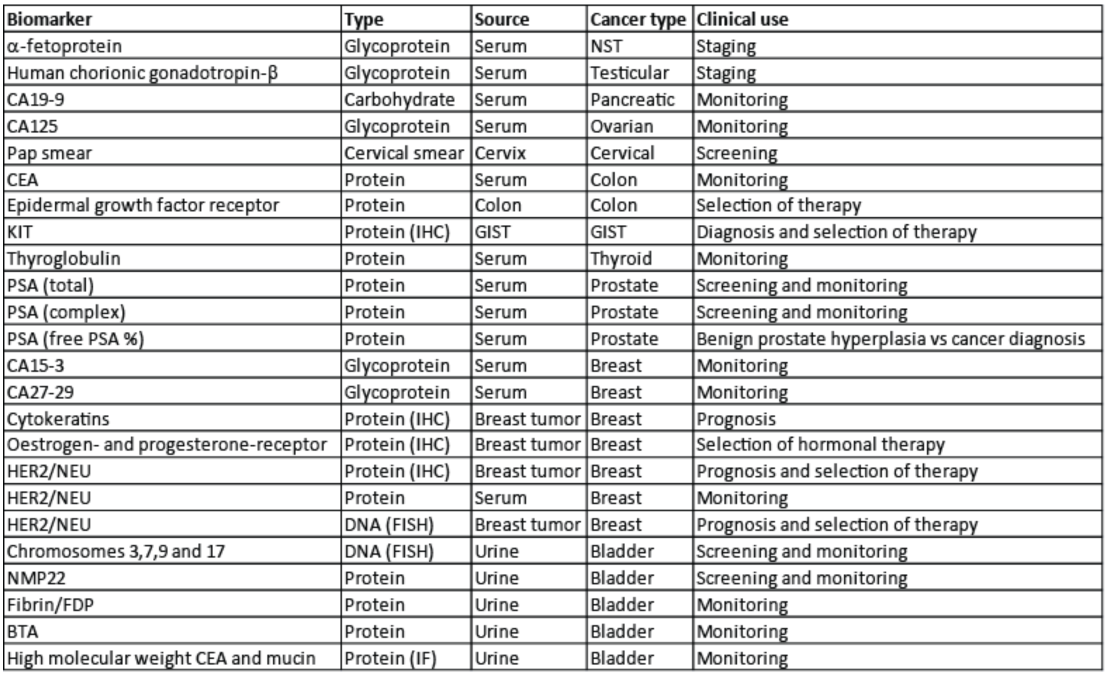
As the focus of this thesis has been method development proteomics studies of clinical materials, the concept of *biomarkers* is of importance, as biomarker discovery often is the end goal when studying clinical materials. To begin with, biomarkers do not have to be proteins. In a broad definition a biomarker could be any molecule, or even an image, used as an indicator of a biological state.

Depending on the purpose of the biomarker there are a few different classes of biomarkers; *diagnostic* markers are used to show presence of a disease, *prognostic* markers on the other hand will tell something about a disease outcome, regardless and independent of drug/therapy. *Prognostic* markers are often confused with *predictive* markers, which can be used to tell how a patient will respond to a treatment. For example, a diagnostic marker can be used to diagnose lung adenocarcinoma. A prognostic marker will indicate that the patient has a good chance of surviving up to three years after diagnosis. A predictive marker will in addition tell that the tumor will not respond to a specific chemotherapeutic drug, but will respond well to radiotherapy. This tailored treatment approach, where every patient will receive the most appropriate medical treatment and the most fitting dosage and combination of drugs based on his or her genetic make-up is called *personalized medicine* [112]. The use of biomarkers is central in personalized medicine as they are needed for therapeutic guidance etc. The concept of personalized medicine, together with technical advances in –omics technologies, has lead to an increased interest in the scientific community for biomarker discovery studies.

1.9.1 Biomarkers in cancer

In cancer therapy, *personalized medicine* is extremely relevant as population based medicine has, in many cancer types, not been successful in curing cancer patients. At present it is very difficult to predict who will respond well to what treatment, as tumors commonly develop resistance to drugs and in addition many patients suffer from severe treatment related side-effects. Besides therapy related markers, *diagnostic markers* are also highly sought after in oncology, as early diagnosis almost inevitably improves the prognosis.

At present only a limited number of biomarkers are approved for use in the clinic for cancer diagnostics, prognostics and therapeutic guidance [113]. See table 1 for an overview of FDA approved biomarkers in cancer.



Biomarker	Type	Source	Cancer type	Clinical use
α-fetoprotein	Glycoprotein	Serum	NST	Staging
Human chorionic gonadotropin-β	Glycoprotein	Serum	Testicular	Staging
CA19-9	Carbohydrate	Serum	Pancreatic	Monitoring
CA125	Glycoprotein	Serum	Ovarian	Monitoring
Pap smear	Cervical smear	Cervix	Cervical	Screening
CEA	Protein	Serum	Colon	Monitoring
Epidermal growth factor receptor	Protein	Colon	Colon	Selection of therapy
KIT	Protein (IHC)	GIST	GIST	Diagnosis and selection of therapy
Thyroglobulin	Protein	Serum	Thyroid	Monitoring
PSA (total)	Protein	Serum	Prostate	Screening and monitoring
PSA (complex)	Protein	Serum	Prostate	Screening and monitoring
PSA (free PSA %)	Protein	Serum	Prostate	Benign prostate hyperplasia vs cancer diagnosis
CA15-3	Glycoprotein	Serum	Breast	Monitoring
CA27-29	Glycoprotein	Serum	Breast	Monitoring
Cytokeratins	Protein (IHC)	Breast tumor	Breast	Prognosis
Oestrogen- and progesterone-receptor	Protein (IHC)	Breast tumor	Breast	Selection of hormonal therapy
HER2/NEU	Protein (IHC)	Breast tumor	Breast	Prognosis and selection of therapy
HER2/NEU	Protein	Serum	Breast	Monitoring
HER2/NEU	DNA (FISH)	Breast tumor	Breast	Prognosis and selection of therapy
Chromosomes 3,7,9 and 17	DNA (FISH)	Urine	Bladder	Screening and monitoring
NMP22	Protein	Urine	Bladder	Screening and monitoring
Fibrin/FDP	Protein	Urine	Bladder	Monitoring
BTA	Protein	Urine	Bladder	Monitoring
High molecular weight CEA and mucin	Protein (IF)	Urine	Bladder	Monitoring

Table 1. FDA approved cancer biomarkers. Modified from [113] GIST=Gastro Intestinal tumors, IHC= immunohistochemistry, IF= immunefluorescence, NST= nonseminomatous testicular

Prostate specific antigen or PSA is probably one of the most well-known biomarkers used in clinical practice today. Although widely used, it is a source of controversy [114] and it illustrates some of the challenges when working with cancer biomarkers. PSA is produced in the epithelial cells of the prostatic glands and is normally only present in very low concentration the in blood stream. In cancer there is an augmented leakage of PSA into the blood stream due to an increased number of epithelial cells, a deficiency in the basal membrane, and because the cells lose their contact with the excretory ducts [115]. A cut-off level of 4ng/ml of PSA in plasma is used to indicate prostate cancer. This illustrates a problem from a biomarkers discovery point of view as the normal total protein concentration in plasma is between 50-100 mg/ml[8] and most proteomics technologies only span over three to four orders of magnitudes in concentration range only reaching proteins in the low µg/ml range (figure 5).

In addition, PSA, as a biomarker, has both a limited sensitivity and specificity. It has been shown that many men diagnosed with prostate cancer have a PSA value below 4ng/ml [116] illustrating the limitation in sensitivity.

The limitation in specificity has several reasons. First, there are several other non-malignant prostate diseases that cause an increase in PSA, such as prostatitis and benign prostatic hyperplasia [117]. Second, other tissues have also been shown to express PSA [118] and PSA is also found among women [119].

Taken the complexity of cancer biology it might not be realistic to expect to find single biomarkers with sufficient sensitivity and specificity, rather a panel of biomarkers might be needed [120, 121]. In 2007 the MammaPrint was approved by the FDA as a prognostic test, used to assess the risk of metastasis in breast cancer [122]. The test is based on a 70 gene microarray and classifies analyzed tumors as low or high risk for recurrence of the disease [123].

1.10 CLINICAL MATERIALS IN CANCER

Clinical samples are commonly used as discovery materials in proteomics studies. The study design of the proteomics experiment will be dictated by the type of material that is investigated and at what time-point/-s the material is collected. The old saying garbage in – garbage out is of particular relevance when studying clinical materials, as the sources of variability is much larger among humans than in model systems such as yeast, cell lines or animal models. Sample variability can be derived from several characteristics such as sample heterogeneity, inter-individual variation, sample handling, preparation differences, etc.

1.10.1 Tumor Tissue

Tumor tissue is an obvious source of biomarkers in cancer, but normal tissue is also of importance, especially as negative control in discovery studies. Using tissue as a starting point, both DNA, RNA and protein can be obtained from the same sample. Tissue can either be obtained fresh, and subsequently frozen directly after a biopsy or surgical resection, or can be fixed in formalin and embedded in paraffin, and stored in pathology archives.

Biopsies have the advantage that they can be used to obtain both normal and tumor tissue, as biopsies are used for diagnostic purposes. However, the sampling is invasive, which means that repetitive sampling is rarely done and furthermore the amount of material obtained is very limited.

Surgical samples contain much more material than biopsies, since surgery is usually performed to radically remove the entire tumor. The tissue adjacent to the tumor can be used as corresponding normal tissue, but it is likely that it is affected by the presence of the tumor. Hence it is recommended to analyze normal tissue that is sampled as far as possible from the resected tumor. Surgical sampling is most commonly only performed once, and after time of diagnosis, which sets the limits as to which studies can be conducted on the material.

To prepare fresh tissue for proteomics studies, the tissue is usually snap-frozen in liquid nitrogen and homogenized by mechanical disruption (ultraturrax or dounce) [124-126] or ultrasonic disruption [127, 128] prior to analysis. By doing this both tumor cells, stromal cells and infiltrated inflammatory cells will be analyzed together. The heterogeneous cell population is obviously a challenge as it can differ between samples, but it is also well recognized that initiation and progression of cancer not only includes the cancer cells, but also the surrounding microenvironment, highlighting the importance to study the tissue as one entity [129].

Formalin fixed paraffin embedded (FFPE) tissue differs from fresh frozen samples as they are chemically modified. The fixation induces protein, as well as nucleic acid, crosslinkage which limits it’s applicability in mass spectrometry based studies. However, there are several studies published on protein and nucleic acid analysis from FFPE [130-132] and FFPE sections can also be analyzed directly by MALDI imaging, where the section is applied to a MALDI target, coated with matrix and analyzed directly in a top-down approach [133, 134].

As tissue sampling involves invasive procedures, it is common to do discovery in these materials, where the concentration of the marker is potentially high, and then try to develop a blood test for selected candidate markers.

1.10.2 Tumor Cells

Tumor cells have the advantage that they comprise a more homogenous sample than total tissue lysates. Tumor cells can either be obtained from non-solid tumors (e.g. leukemia [135]), from fluids (e.g. blood [136], bronchoalveolar lavage [137], or fine needle aspirates [138]) or prepared from tissue samples using laser micro dissection [139]. As a rule, tumor cell suspensions contain less material than tissue preparations, which could be a challenge, however the potential advantage would be that the markers are enriched in this population of cells. In addition, a selected population of cells, for example cancer stem cells, can be specifically enriched and targeted in the analysis. Obtaining 'normal' cells is equally challenging as in tissue proteomics and similarly repetitive sampling is rarely possible.

1.10.3 Plasma

Plasma is the liquid component of blood and makes up about 55% of the total blood volume. Plasma contains mostly water (90%), but also proteins, glucose, metabolites etc. Plasma is prepared from blood through centrifugation, where the cells are separated from the fluid. If the tube contains anti-coagulants the fluid is defined as plasma, as opposed to serum where the blood is allowed to coagulate and the clot is separated together with the cells.

Plasma is an ideal source of biomarkers from a clinical point of view; the sampling in minimally invasive, repetitive sampling is possible and the sampling is routinely performed in the clinic. From a biological perspective plasma is also a promising source of biomarkers at it is in contact with all organs and tissues, and therefore potentially could contain trace markers from all biological processes in the body.

As plasma only contains very little DNA and RNA much hope has been put into proteomics based discovery of clinically useful biomarkers from plasma.

In 2003 the human plasma proteome project (HPPP) was launched within the human proteome organization (HUPO). HPPP had three major objectives; (1) comprehensive analysis of the protein constituents of human plasma and serum; (2) identification of physiological, pathological and pharmacological sources of variation within individuals over time, leading to validated biomarkers; and (3) determination of variation across individuals and across populations due to genetic, nutritional, lifestyle and other factors [140, 141]. Despite major efforts, none of these goals have been reached. This is due to several specific analytical challenges related to plasma biomarker discovery.

First, plasma has a very high dynamic range of protein concentrations, spanning over at least 10 orders of magnitude [8]. This wide range of concentrations cannot be covered by proteomics technologies, as touched upon in the introduction. However, this would be of less importance if the potential markers where present in high concentrations. This is not the case though, as the classical plasma proteins that exert their function in plasma are highly abundant, in contrast to the low abundant tissue leakage markers that could potentially be used as biomarkers (figure 5). Further, as the markers have no function in plasma they are most likely present in plasma for a limited time-span, before they are degraded.

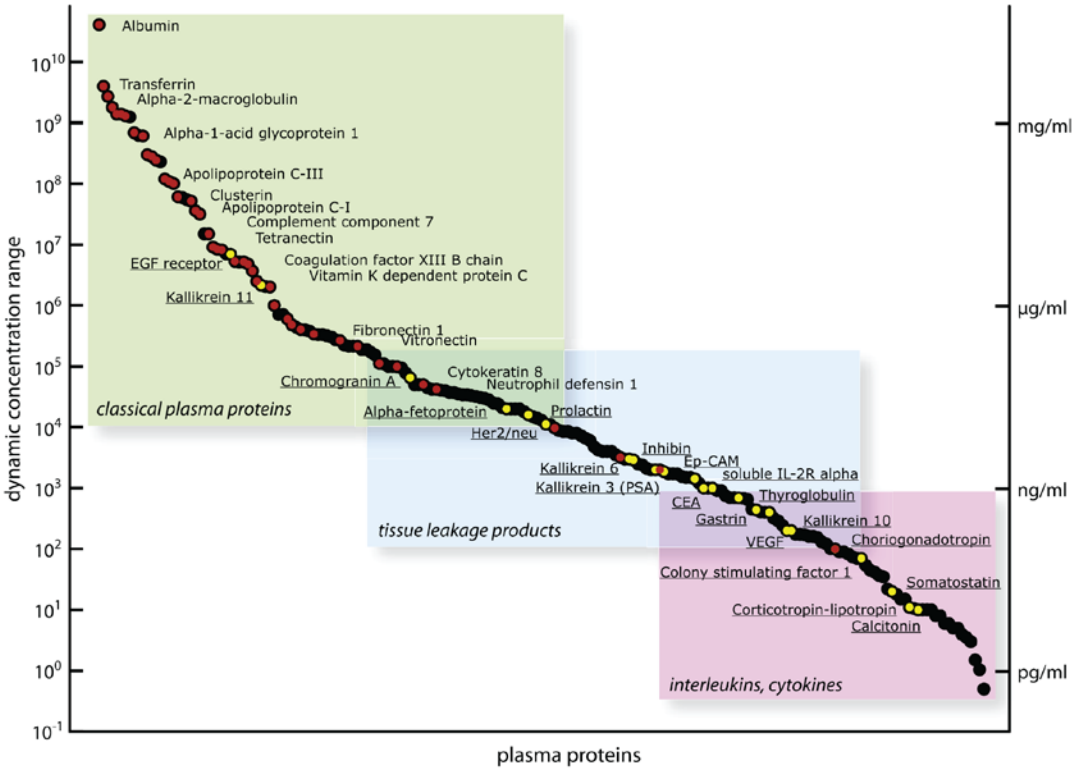


Figure 5. Plasma protein concentrations as depicted in[8]. The proteins are grouped into three main categories; classical plasma proteins, tissue leakage products and interleukins/cytokines. Red dots indicate proteins that have been identified by the HUPO plasma proteome initiative[142] and yellow dots represent currently used biomarkers. Picture adapted from[143] with permission from the publisher.

Taken together these analytical challenges have led to a shift where few discovery studies are performed in plasma, instead discovery is performed in other materials with potentially higher concentration of the marker, and then the validation phase is performed in plasma [143, 144].

1.10.4 Proximal fluids

Proximal fluids are a group of pathological and normal biological fluids that are found in a limited space in the body. The potential advantage with proximal fluids is that they are closer to the organ of interest, and therefore might contain a higher concentration of the marker, which is of advantage in particular for discovery proteomics. Since the marker is released into a fluid, the likelihood that it will end up in plasma might also be higher. Proximal fluids include (among others); cerebrospinal fluid (CSF) – which surrounds the central nervous system [145, 146], bile – which is produced in the liver and stored in the gallbladder [147, 148], amniotic fluid - which fills the amniotic sack in pregnant women [149, 150], saliva – which is present in the oral cavity [151, 152], synovial fluid – which lubricates the joints [153, 154], tear fluid – which is excreted from the eye [155, 156] and nipple aspirate fluid which is derived from the nipple [157, 158] or pathological fluids such as pleural effusion from the pleural cavity [159, 160].

A challenge with proximal fluids is that they are often similar to plasma regarding protein content and the high dynamic range of protein concentrations.

2 THE PRESENT STUDY

2.1 AIMS

The general aim of this thesis was to evaluate and optimize the different stages in mass spectrometry based biomarker discovery from clinical material.

The specific aims were:

Paper I: To develop an analytical workflow for selection of candidate biomarkers from SELDI-MS data.

Paper II: To evaluate three protein prefractionation methods for mass spectrometry based plasma proteomics.

Paper III: To review and evaluate the analytical depth among affinity prefractionation methods for mass spectrometry based plasma proteomics

Paper IV: To explore the possibility of using narrow range iso-electric focusing as prefractionation method of plasma and pleural effusion prior to mass spectrometry based proteomics.

Paper V: To optimize a protocol for tumor cell enrichment for mass spectrometry based proteomics.

2.2 MATERIAL AND METHODS

The materials and methods used in **paper I-V** are described in detail in each paper, and will not be presented meticulously in this section. Instead methodological considerations will be discussed, together with a brief presentation of the purpose of using the method.

2.2.1 General description of the KBC biobank

At Karolinska Biomics Center (KBC) we are currently collecting several different types of clinical materials, with a focus on samples related to lung cancer. The collections are approved by the ethics committee at Karolinska Institutet and are being conducted within the section for thoracic malignancies of the Karolinska University hospital Biobank. All patients have signed an informed consent.

The plasma biobank was set up in 2004 and currently contains approximately 1600 samples. All patients that are assigned to bronchoscopy at the Outpatient division at the department of Respiratory medicine and allergy at Karolinska University Hospital are asked to donate blood and the collection therefore both includes malignant and non-malignant samples.

Pleural effusion has been collected since 2005 and the biobank consists of 100 samples. All patients who have pleural effusion drawn at the thorax clinic are asked to participate in the study, and at the same time as the pleural effusion is removed blood is also collected.

Tissue samples, as well as plasma samples, are obtained from all patients that go through surgery due to suspected lung cancer and have signed informed consent. At present 130 samples have been collected since the start in 2006.

2.2.2 Plasma and pleural effusion

In **paper II** and **paper IV** plasma and pleural effusion is analyzed. Both samples have been prepared using a standard operation protocol (SOP) that has been developed in-house. The SOP includes both preparatory considerations as well as data collection. EDTA tubes was chosen as collection tubes after an initial protein degradation study, which showed less protein degradation in EDTA plasma over time, compared with serum, heparin plasma, citrate plasma and gel plasma (unpublished data). The EDTA tubes were also routinely used in the clinic, which facilitated the logistics of the collection, and were recommended by HPPP [161]. In parallel with the sample collection, data is also gathered on the time of sampling, time of sample preparation, level of hemolysis and in addition, clinical data and data from clinical chemistry analysis (kemlab) is collected to ensure high quality of the selected samples.

2.2.3 Lung cancer tumor tissue

In **paper V** a method for preparation of tumor cell suspension from lung cancer tissue is described. A SOP has been developed both for the sample preparation of tumor tissue as well as the data collection. A technician from our lab collects the surgical specimen directly after it has been removed. The tumor tissue is cut and one piece is snap-frozen, and in parallel, one piece is prepared into a cell suspension. Cytospin as well as a tumor imprint is prepared for quality control. Adjacent normal tissue is prepared according to the same protocol as for tumor tissue. In addition, archived formalin fixed paraffin embedded sample is prepared and stored in the biobank. The tumor database and the plasma and pleural effusion databases are connected so that information on sample availability and clinical data is easily accessed.

2.2.4 Acute myeloid leukemia cells

The acute myeloid leukemia (AML) cells analyzed in **paper I** where obtained at time of diagnosis from peripheral blood. One of the challenges in this study was the limited amount of material and therefore leukemic cell lines where analyzed in parallel to investigate the potential of using these model systems in future follow-up studies. As AML is a very heterogeneous disease all samples were evaluated and scored by a pathologist for a second diagnostic evaluation and approximation of cell content.

2.2.5 High abundant protein depletion

Alongside the Multiple affinity removal system (MARS) column (Agilent technologies) used in **paper II** and **IV** several other depletion systems were evaluated, (primarily based on reproducibility and compatibility with downstream analysis) before settling with the MARS-7 column. The MARS-7 column is specifically designed for plasma rather than serum as it, in addition to albumin, IgM, IgA, transferrin, antitrypsin, and haptoglobulin, also removes fibrinogen – present only in plasma. The column is available both as a spin column and a LC-column, and the LC-column was chosen because of its' high sample capacity, the increase in throughput and the potential reduction of variability by coupling to an automated FPLC system.

In addition to plasma and pleural effusion, we have also used MARS columns to successfully deplete CSF and synovial fluid from high abundant proteins, showing the robustness and versatility of the system (unpublished data).

2.2.6 iTRAQ labeling

The iTRAQ label has been used for quantification in both **paper II**, **IV** and **V**. At present eight different isobaric labels are available. This means that up to eight samples can be pooled analyzed as one. The iTRAQ label is primarily used for relative quantification, and the ratio between the reporter ions within one spectrum is used for quantification of each peptide within one pooled sample (figure 6).

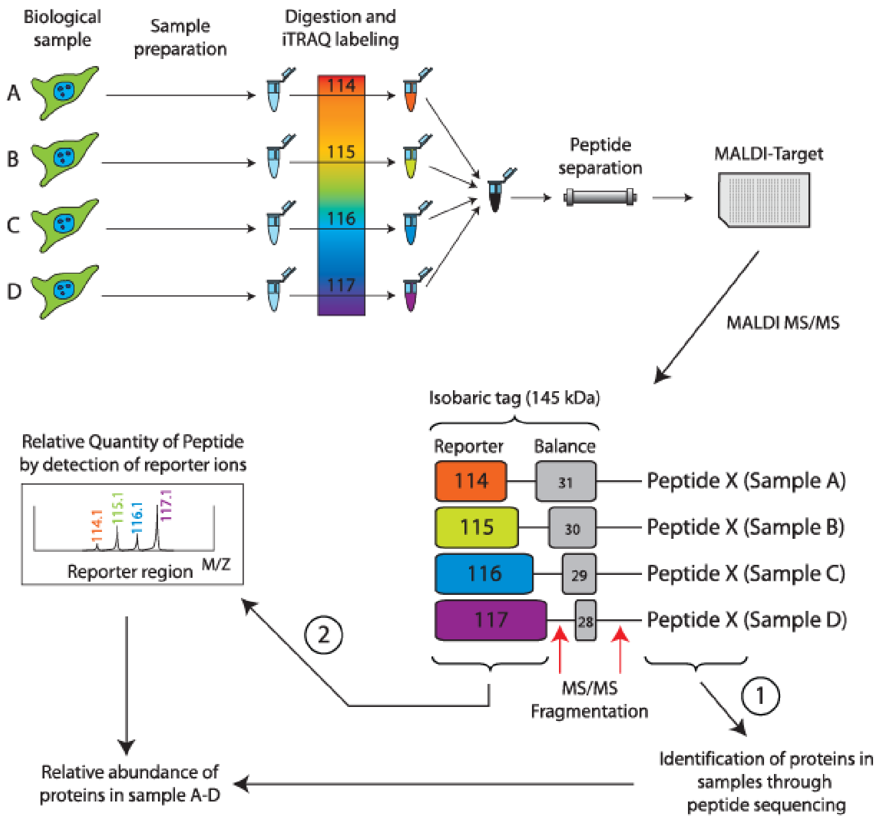


Figure 6. Basic principle of the iTRAQ labeling technology. In this example four samples are labeled and analyzed using LC-MALDI-TOF/TOF. Courtesy of Lukas Orre.

If one wants to include more than eight samples in one experiment comparison *between* pooled samples is necessary, instead of comparisons only *within* one pooled sample. To enable this one can use an internal standard that is shared between the pooled samples. As the standard needs to be present in all spectra it needs to cover all peptides present in the sample. The easiest way to construct such a standard is by pooling the individual samples in the study to one pooled internal standard, as performed in **paper IV**. The pooled internal standard is then included in all individual 8-plex experiments. The pooling of the internal standard is preferably performed on the peptide level, to ensure that all peptides present in the individual samples are present in the internal standard. When applying the pooled internal standard approach, a few characteristics of the iTRAQ labeling become evident. First, different peptides and proteins are identified in iTRAQ samples that are analyzed separately. I.e. the proteins identified from the pooled internal standards are not the same in the individual pooled samples. Second, if a peptide is identified in one of the samples within a pool it is also identified in all the other samples within that pool. Third, quantitative differences within one pool rarely exceed 20%. The two latter observations could be derived from the fact that the iTRAQ reporter ion ionizes very well, and that the dynamic range of the mass spectrometer is limited, thereby quenching strong signals and generating and over-estimation of low-intensity signals[162]. This is of course of importance both when designing an iTRAQ experiment and when analyzing the data.

2.2.7 Narrow range peptide isoelectric focusing

The rationale behind using narrow range peptide isoelectric focusing is to reduce the complexity induced by tryptic digestion, by selectively analyze a sub-fraction of peptides with an acidic p*I*. The p*I* range was chosen as it has previously been shown that at least 80% of human proteins have at least one tryptic peptide between pH 3.5-4.5 [71, 73]. By analyzing this sub-fraction of peptides the complexity of the sample can be reduced without significant loss of proteome coverage (figure 7). As the theoretical p*I* of peptides can be calculated, the p*I* of the identified peptides can be used to validate the peptide sequence (identified peptides with p*I* outside the pH range 3.5-4.5 are more likely to be false positives). In addition, this approach is compatible with iTRAQ labeling as the different iTRAQ labels migrate similarly in IEF [75].

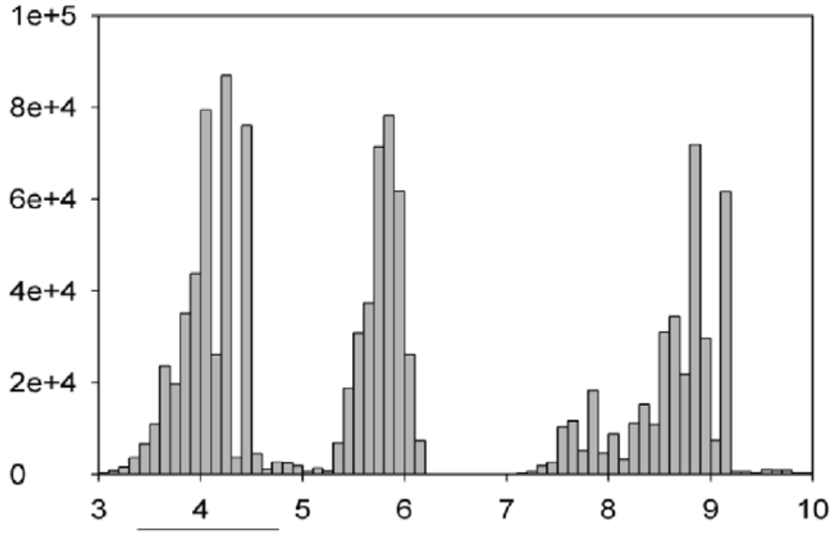


Figure 7. A plot of the predicted p*I* values for human tryptic peptides. All peptides with 4-60 amino acids and no missed cleavages are included. Approximately one third is in the pH interval 3.5-4.5, indicated by a black bar. Courtesy of Hanna Eriksson.

In **paper IV** and **V** free flow electrophoresis (FFE) and immobilized pH strips was used for narrow range peptide isoelectric focusing. The FFE system has the advantage that it performs the separation in solution, which is directly compatible with downstream LC-MS/MS analysis. Using the IPG strips, the strips have to be either manually cut, or eluted using a robot, which is not commercially available today [73]. The manual cutting has its’ drawbacks, as it relies on a steady hand that can cut pieces of even width and with a 90° angle so that the fractions become equally wide. Using the cutting strategy it is preferred to analyze continuous fractions to reduce strip to strip variation.

Another technology where the peptides are separated both in gel and in solution is the OFFgel technology, where the peptides can be directly obtained from the solution without an elution step. At present there is no strip available for the OFFgel system for narrow range IEF in the pH range 3.5-4.5. To evaluate OFFgel's potential for separation in this pH range we, in our lab, tried to separate peptides on 3.5-4.5 strips from GE-healthcare using the OFFgel system. This approach proved to be less applicable on the OFFgel system as the majority of the strip dried out and all fluid was contained in the most basic fractions. Most probably there was an osmotic counter flow of the fluid trying to equalize the difference peptide concentration over the gradient, as the majority of peptides would fall outside the 3.5-4.5 p*I* range and therefore end up in the most basic end of the strip (unpublished data).

In **paper V** a custom made strip optimized for narrow range peptide isoelectric focusing was used, optimized to generate less background in mass spectrometry, and made up by a custom made gradient (pH 3.7-4.9) to target as many proteins as possible.

2.2.8 LC-MALDI-TOF/TOF

The LC-MS/MS set-up used in **paper II, IV** and **V** was an off line nanoLC system (dionex) coupled to a MALDI spotter. The samples were subsequently analyzed using an ABI 4800 MALDI-TOF/TOF. NanoLC is a LC technique using columns with an internal diameter between 10-150 µm. The name nanoLC refers to the mobile-phase flow rate which is in the nanoliter per minute range. The main advantage of using smaller columns is the increased detection sensitivity and the improved separation (higher resolution) that can be obtained as a result of reduced sample dilution and decreased particle sizes in the columns. However, when reducing the particle size the column pressure increases, as a result of reduced interstitial void between the particles. In **paper II** a standard reversed phase C18 column was used; in **paper IV** and **V** a monolithic column was used. Instead of a carbon chain coupled to spherical particles as in traditional reversed phase, the monolithic column is made up by a continuous network, resulting in lower column back pressure, and thereby enabling higher flow-rates and shorter gradient times. The continuous network results in no interstitial void of in column, which reduces the diffusion of the analytes and increases the resolution of the separation.

The MALDI-TOF/TOF mass spectrometry analysis of the samples enabled identification of the peptides and further has good compatibility with iTRAQ labeling and quantification.

2.2.9 SELDI-TOF

Used in both **paper I** and **II**, SELDI was the main top-down approach applied in this thesis. SELDI was first described by Hutchens and Yip in 1993 [43] and is a high throughput chip based MALDI technique where the sample is analyzed directly on a selective solid-phase affinity surface. In addition to reduction of complexity, the chromatographic surface allows for concentration and washing of the sample, which facilitates the mass spectrometry analysis of biological samples. Antibodies can also be coupled to the SELDI chips and this was used for immuno-capture of S100A6 in [163, 164]. SELDI analysis is biased for analysis of low molecular weight proteins and peptides (<30kD) as the ionization is most effective in this mass range.

2.2.10 Tissue microarray

Tissue micro array, used in **paper IV** to validate potential markers, enables high-throughput immunohistochemical (IHC) analysis of formalin fixed paraffin embedded samples [104]. As the samples are evaluated by a pathologist it is important that the pathologist is not biased and does not know anything about the underlying study question.

2.3 RESULTS AND DISCUSSION

2.3.1 Paper I

Proteomic Data Analysis Workflow for Discovery of Candidate Biomarker Peaks Predictive of Clinical Outcome for Patients with Acute Myeloid Leukemia

The background of this paper was that we had previously worked with the SELDI technology in our group using standard protocols [164] and we had in parallel started working with optimizing bottlenecks in the SELDI approach. A novel peak detection algorithm was developed [165, 166] and in this paper we wanted to present an optimized analytical workflow for SELDI experiments including; a) clinical sample selection, b) experimental optimization, c) repeatability estimation, d) data preprocessing, e) data fusion and f) marker selection. The clinical question at hand further motivated this development as the clinical material was very limited in amount, heterogeneous both in regard of cell content and clinical data, as well as presented a skewed distribution of the clinically most important outcome parameter. The clinical aim of the study was to identify markers that would predict the duration of complete remission (CR) among patients diagnosed with acute myeloid leukemia (AML). This information would be of prognostic value and assist in therapeutic decisions, as CR is directly related to overall survival in AML.

The clinical material consisted of blast cells from peripheral blood from patients diagnosed with AML, sampled at the time of diagnosis. Out of the approximately 200 samples available only 58 were selected for SELDI analysis, as they contained enough cells. Out of the 58 samples some patients were later shown to be non AML and in addition the vast majority of the patients had a very short CR. To create two separate classes for patients with long and short complete CR durations two extreme groups containing only the first (n=13) and the fourth quartiles (n=13) of the CR were created.

Since the amount of clinical material was limited, two leukemic cell lines, KG1 and NB4 were included in the study, to evaluate their potential as sources of biomarkers in future biomarker purification steps.

To obtain as much information as possible from the SELDI MS analysis, an initial optimization was performed. Three different chromatographic chip surfaces (weak cation exchange CM10, strong anion exchange Q10 and reversed phase) were evaluated in combination with three different types of matrix and three buffer compositions (unique for each surface). Once two chipsurfaces (C10 and Q10) with optimal buffer conditions and matrix had been chosen (based on the number of detected peaks and signal-to-noise ratio) a reproducibility test was performed, showing overall CVs below 20%.

SELDI-MS data processing generally includes mass calibration, spectral smoothing, baseline correction, normalization and peak detection and alignment. As we had previously identified the peak detection and alignment as a bottle necks with a large need of manual inspection, four methods for peak detection and alignment were evaluated; annotated regions of significance (ARS) developed in-house [165, 166], Ciphergen express which is the standard software for SELDI-MS data, segment wise spectral alignment (PAGA) followed by binning, and direct binning of the raw data.

To take advantage of the additional value of using two chipsurfaces and select the most promising markers from the two experimental setups, the individual results from the SELDI-MS were fused. A standard SELDI uni-variate approach using Mann-Whitney t-test in combination with the standard Ciphergen express peak detection method was compared to a multivariate hierarchical orthogonal partial least squares (O-PLS) analysis approach. The number of variables picked by the O-PLS and the sensitivity and specificity of each approach can be found in table 2.

	precision	Sensitivity	specificity	n o var in model
ARS[a]	0.91	0.79	0.91	64
CE[b]	0.81	0.73	0.82	47
PAGA[c]	0.77	0.72	0.76	380
binned[d]	0.82	0.74	0.83	387
CE-p[e]	0.78	0.82	0.74	50

[a]Annotated regions of significance, [b]Ciphergen Express, [c]Peak alignment by a genetic algorithm + subsequent binning (6-fold), [d]Binning (6-fold), [e]A multivariate model based on the 50 most significant ($p < 0.07$) peaks from Ciphergen Express –“the standard method”.

Table 2. Precision, sensitivity and specificity of the biomarker selection models based on the different peak detection and alignment methods. n o var in model = number of variables in the model. Variables can be either peaks or m/z values depending on the peak detection and alignment method used.

From all significant features, 21 protein peaks were chosen as the most promising markers, as they were selected in several of the different approaches.

The main conclusion from this study is that the data output will benefit from using several methodologies, both in the data pre-processing, and in the biomarker selection. As any limited set of clinical samples will be biased (because a small number of samples rarely cover all the normal biological variation between individuals), a data analysis of protein profiles from a limited set of samples will also be biased. For this reason, it will be impossible to find the “optimal” data pre-processing method based on a limited cohort. However, by establishing the final biomarker selection from different data analysis methodologies, the findings are more likely to be robust. The workflow presented in this paper is not exclusive to the SELDI technology, and could easily be adapted to other label-free quantification experiments, using for example a MALDI or LC-MS/MS platform.

2.3.2 Paper II

Evaluation of Three Principally Different Intact Protein Prefractionation Methods for Plasma Biomarker Discovery

As we had started setting up our plasma biobank we wanted to determine how we should prepare the plasma in order to increase the likelihood of detecting clinical relevant biomarkers. Primarily we wanted to evaluate intact protein prefractionation as a first line separation approach, since the intact protein approach had the benefit that it could be combined with separations on the peptide level and be used together with several analytical techniques. Further, we wanted to limit the number of fractions generated, as the discovery phase in proteomics is both labor intense, time-consuming and expensive, and therefore only a limited number of samples are usually analyzed. Three prefractionation methods were evaluated; high abundant protein depletion, ProteoMiner beads and an in-house developed size fractionation method. Using SELDI-MS, LC-MALDI-MS/MS and SDS-PAGE we wanted to investigate the individual method's compatibility with downstream analysis, reproducibility and analytical depth.

High abundant protein depletion was performed using the MARS-7 (Agilent technologies) LC column, targeting albumin, transferrin, IgG, IgA, antitrypsin, haptoglobulin and fibrinogen. Both the depleted plasma and the fraction containing the removed proteins were analyzed.

The ProteoMiner beads (BioRad) is a novel affinity based prefractionation method with a combinatorial peptide ligand library coupled to beads. The technique takes advantage of the fact that the individual peptides have different binding properties, and as the beads contain equivalent binding capacity (identical number of peptides), the beads binding high abundant proteins will be saturated, whereas there will be a potential concentration of low-abundance proteins. The beads are then eluted with four different eluents, thereby generating four fractions.

The size fractionation method developed for this evaluation aimed at targeting both low molecular weight proteins in solution in plasma and low molecular weight proteins bound to larger proteins. The rationale behind this was to take advantage of the fact that most classical plasma proteins are larger than 50kD to be able to exert their function in plasma, as approximately 50kD is the kidney filtration limit. Further it is likely that tissue leakage proteins in solution might only be present in the blood stream for a limited time, whereas proteins bound to carrier proteins in plasma might present in the blood for a longer time.

Compatibility with downstream analysis was evaluated using SDS-PAGE (gel based intact protein analysis), SELDI-MS (mass spectrometry based intact protein analysis) and LC-MALDI-TOF/TOF (mass spectrometry based peptide analysis).

Overall all the three methods showed good compatibility with the SDS-PAGE.

In the SELDI-MS analysis the fractions from the MARS-7 column showed few peaks and high variability, and would not be the method of choice. ProteoMiner on the other hand performed very well on the SELDI platform, with a high number of peaks and a high reproducibility.

In the LC-MALDI-MS/MS section two different approaches were evaluated, both direct digestion of the fractions, with no second line of fractionation, and a GeLC-MS/MS approach where the samples first were separated on SDS-PAGE and sections of the gel subsequently cut out and digested. The two different shotgun approaches highlighted the advantage of a second line of fractionation, as approximately twice as many proteins were indentified after the GeLC-MS/MS as compared with the direct MS/MS approach. When adding up the number of identified proteins from each fractionation method, all prefractionation methods identified more proteins than were identified from the crude plasma reference sample. Comparing the individual fractions, the highest number of identified proteins was obtained from the size fractionation (n=123) followed by the MARS-7 (n=116). However when adding the identities from the individual fractions together the ProteoMiner beads gave the most identified proteins (n=150), not very surprising as the ProteoMiner generates twice as many fractions as the MARS-7 and the size fractionation.

The reproducibility of the methods was calculated primarily using SELDI peak intensity, but for the high abundant protein depletion and the size fractionation iTRAQ reporter ion intensities were also used. The ProteoMiner beads proved to be the overall most reproducible of the methods based on the SELDI data. The size fractionation and the MARS-7 column showed mixed results with good reproducibility in one of the fractions and (*low* and *flowthrough*) and poor reproducibility in the other fraction (*cut off* and *eluate*).

In general, analyzing analytical depth is quite difficult in proteomics experiments. The number of identified proteins can be used as a measure of how large fraction of the proteome that is identified, but when analyzing plasma one rarely wants to cover the plasma proteome, but rather identify a large number of tissue leakage proteins. As the identification of low abundant proteins is hindered by the presence of high abundant proteins, the aim is to avoid repetitive identification of classical plasma proteins. Therefore, in this study, the analytical depth was evaluated by looking at the distribution of Ingenuity pathway analysis terms (Ingenuity systems) related to cellular compartment among the identified proteins. In this analysis the extracellular proteins completely dominate all samples, and this is further confirmed when analyzing the pathways that are represented among the identified proteins. The top pathways in all samples are, with no exception, related to functions inherent to plasma, such as coagulation, acute phase reaction and complement cascade.

The three methods were all based on different separation principles, high abundant protein removal, protein size separation and equalization of protein concentration.

Some general observations about the methods performance could be noted. First, it is obvious from both the SDS-PAGE and the MS/MS data that the high abundant protein depletion removes more proteins than aimed for. More than 30 proteins are identified from the *eluate* in this experiment. These proteins could either be unspecifically bound to the column, or bound to the proteins that are depleted. In addition we analyzed the sequence coverage of two of the high abundant proteins targeted by the MARS-7 column (albumin and fibrinogen). Protein sequence coverage was used as surrogate marker for protein concentration. (A high sequence coverage indicating a high protein concentration, and a low sequence coverage indicating a low protein concentration). This analysis showed a very high efficiency of the albumin removal, as no albumin is present in the flowthrough. Fibrinogen on the other hand could be found in both fractions, with an even higher estimated concentration in the flowthrough. This could be due to protein cleavage where the targeted epitope is removed.

The ProteoMiner beads aim at reducing the protein concentration differences in plasma, and this can be seen both on the gel and on the protein concentration analysis. There is, however, a quite large overlap in identified proteins between the four fractions, so pooling the fractions pair-wise could be useful, as it would reduce the numbers of fractions to be analyzed, but not influence the analytical depth noteworthy. The size fractionation proved to be rather inexact in its molecular weight separation leading to presence of high molecular weight proteins in both fractions. Further, the cut off filters induced a large variability. The potential of this type of separation was highlighted when comparing the GeLC-MS/MS experiment with a direct LC-MS/MS. Once the high molecular weight proteins were analyzed separately, the number of identified protein more than doubled.

Taken all these results together we decided to use the MARS-7 depletion column in our plasma studies and primarily analyze the *flowthrough* fraction only. The generation of only one fraction is a clear advantage, together with the high number of identified proteins both from the direct LC-MS/MS and the GeLC-MS/MS approach, the high reproducibility, and the high level of automation.

2.3.3 Paper III

Affinity prefractionation for MS-based plasma proteomics

After having evaluated the three prefractionation methods in **paper II** and seen that the increase in proteome coverage with a higher number of identified proteins mostly was reflected by an increased detection of classical plasma proteins we wanted to study affinity enrichment approaches for low abundant proteins as a second line of fractionation. A literature study was performed to evaluate the currently available affinity prefractionation technologies for plasma mass spectrometry based plasma proteomics. Both intact protein fractionation and peptide fractionation were reviewed. The main focus of the review was global proteomics analysis, but we also wanted to investigate affinity enrichment for targeted proteomics technologies such as selected reaction monitoring (SRM). The affinity methods evaluated include; high abundant protein depletion, ProteoMiner beads, carbonylated protein enrichment, cystein containing peptide enrichment, lectin affinity enrichment, hydrazide chemistry enrichment and metal affinity enrichment. To evaluate the analytical depth of the individual enrichment methods, a meta-analysis was performed using public domain data. A similar analytical approach as in **paper II** was adapted, the supplementary lists with identified proteins from the individual experiments were downloaded and the distribution of protein localizations was analyzed to investigate a potential enrichment of tissue leakage proteins. Interestingly, when we started to analyze the data there seemed to be a stronger relationship between the number of identified proteins and the analytical depth, rather than a method dependency.

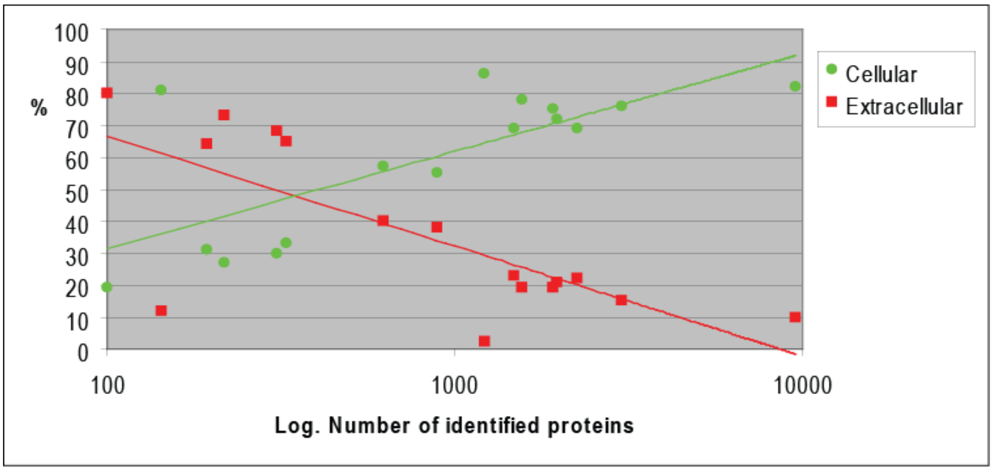


Figure 8. Correlation between cellular and extracellular proteins and number of indentified proteins, based on 16 different affinity enrichment data sets. Each data set has been divided into two sub-classes; cellular proteins (green) and extracellular proteins (red). Percent of total number of proteins in each data set belonging to the cellular or extracellular sub-class (Y) is plotted against total number of proteins in the data set (X).

This was in line with the results from **paper II** were approximately 100-150 proteins were identified and subsequently were mostly assigned to the extracellular space. The results from this meta-analysis further pointed us towards the benefits of extensive fractionation. However, it also highlighted the challenges of in-depth proteomics analysis of larger sample sets of plasma samples, as the throughput is negatively affected by extensive prefractionation.

2.3.4 Paper IV

Use of narrow-range peptide IEF to improve detection of lung adenocarcinoma markers in plasma and pleural effusion

Paper II and **paper III** highlighted the potential of extensive fractionation to increase analytical depth in plasma biomarker discovery studies. Previous studies from our lab had shown high number of identified proteins and good reproducibility of narrow range isoelectric focusing (IEF) on the peptide level on cell line material [73, 75]. The separation was performed in the pH range between 3.7-4.9.

In this paper we wanted to evaluate this methodology on plasma, in combination with high abundant protein depletion. In the previous studies IPG strips had been used for the IEF, but we also wanted to evaluate free flow electrophoresis (FFE) as a possible platform for the separation. In FFE the separation is performed in solution, and the peptides are fractionated directly into a 96-well plate, which is appealing as it removes the elution step performed when using the IPG strips. All samples were therefore separated both with the FFE and the IPG strips and subsequently analyzed using an iTRAQ-LC-MALDI-TOF/TOF approach.

To assess the clinical applicability of the workflow three samples from patients diagnosed with lung adenocarcinoma and three patients diagnosed with pleurits where chosen. The clinical aim was to discover markers correlated to presence of malignancy. In addition to plasma we also wanted to explore pleural effusion, a proximal fluid in lung cancer, and its' applicability for biomarker discovery in mass spectrometry based biomarker discovery. Both plasma and pleural effusion were obtained from all patients, and treated similarly all through the analytical workflow.

In the previously published cell line study approximately 3700 proteins had been identified using this methodology, so we were quite disappointed to see only approximately 100-300 proteins to be identified from the plasma and the pleural effusion using the two different IEF technologies. Approximately twice as many proteins were identified using the IPG strips as compared with the FFE. This difference is probably a reflection of the difference focus precision, where approximately 80% of all peptides could be found in one or two fractions in the IPG experiment, compared with only 50% in the FFE.

However, this does not explain the difference between the previous cell line experiment and the current plasma and pleural effusion experiment, as the spread over fractions is similar in the two studies. After studying the performance of the IEF step and the liquid chromatography step, there seemed to be no apparent difference in separation performance between the two studies. Surprisingly, when comparing the number of significantly identified peptides, the number was approximately in the same order, 8157 peptides from one sample in the cell line experiment and 6153 peptides from one sample in the plasma experiment, despite the large difference in number of identified proteins. Further, the number of peptides successfully assigned to a protein was also similar, showing good performance of both MS/MS analyses. However, when comparing the sequence coverage among the proteins an obvious difference emerged. The sequence coverage among the identified proteins differed significantly, 13% of the proteins from the plasma experiment had a sequence coverage below 10%, whereas the corresponding number from the cell line experiment was 76% below 10% sequence coverage. Again, this illustrates the problematic nature of working with plasma as a discovery material. Our hypothesis was that pleural effusion would comprise an intermediate fluid between the tumor and the plasma, and therefore contain an enrichment of tissue leakage markers. The plasma proteome and the pleural effusion proteome had previously not been compared, but in, analogy to the comparison above, looking at the similarity in number of proteins identified alone (282 vs 300), our guess was that they would be quite similar in protein content as well. As expected, about two thirds of the proteins overlap, and in addition there was no difference in the types of proteins identified as well. Much in line with the results from **paper II** and **III**, the majority of the proteins could be assigned to the extracellular space.

However, going back to the clinical question, more proteins were found to be significantly differently expressed when trying to detect markers for presence of lung adenocarcinoma in pleural effusion than in plasma.

To choose markers for initial validation the first step was to combine the results from the different analytical approaches (FFE, IPG, plasma, pleural effusion) and, as in **paper I** select some markers that were common over several of the analytical conditions. Based on this, antibody availability and published data, nine markers were chosen for initial validation. One of the nine potential markers (NPC2) that showed a strong up-regulation in pleural effusion was chosen to test the hypothesis that the markers could be derived from the tumor. Formalin fixed paraffin embedded tissue from normal lung as well as tissue from lung adenocarcinoma was stained for NPC2. IHC showed low to moderate staining in normal lung and high staining in the samples from lung adenocarcinoma, thereby emphasizing the potential of finding tissue derived markers in proximal fluids.

As the end-goal of a plasma discovery experiment would be a plasma based test we wanted to do the validation in crude plasma. Out of the nine potential markers, the expression pattern of four of the markers (A2M, SERPINA 1, EFEMP1 and CLEC3B) could be validated in crude plasma using western blot.

In summary **paper IV** highlights the potential of using proximal fluids for biomarker discovery and shows that narrow range IEF can be applied to a plasma and pleural effusion proteomics workflow. Due to the protein composition differences between cells and body fluids, additional fractionation on the protein level, or improved resolution and higher dynamic range of the mass spectrometry analysis, is probably needed to expand the number of identified proteins.

2.3.5 Paper V

A novel method for sample preparation of fresh lung cancer tumor tissue for proteomics analysis by tumor cell enrichment and removal of blood contaminants

Paper IV showed high similarity between plasma and pleural effusion and limited benefit of narrow range peptide isoelectric focusing on these body fluids. Introducing additional fractionation steps to increase the proteome coverage in plasma and pleural effusion would introduce variability and further reduce the throughput in the discovery phase. This pointed us towards using tumor tissue as discovery material and plasma and pleural effusion as validation material.

As tissue heterogeneity and blood contamination are challenges when working with tumor tissue we wanted to explore the possibilities of preparing an enriched tumor cell suspension. By mechanical mincing, erythrocyte lysis, sample-wash and cell filtration we aimed primarily at removing blood contaminants and secondarily at removing stromal components. Eight tissue samples were chosen for evaluation of the method, six lung tumors (two adenocarcinoma, two squamous cell carcinoma, two large cell carcinomas) and two normal lung samples. As reference, histological sections of the tissues were stained in parallel to the tissue preparations. Direct lysis of snap-frozen corresponding tissue was used as standard protocol control. To analyze the cell content of the samples, cytospin glasses were prepared from all samples. Reproducibility of the novel method as well as analytical depth was evaluated using an iTRAQ-LC-MALDI-TOF/TOF workflow.

Cytological analysis of the samples showed that the percentage of tumoral cells ranged from 20 to 70 percent in the enriched tumor suspension (ETS) samples. The cytospins from the normal tissue contained mostly inflammatory cells and as little as 10% epithelial cells. The other dominating cell type was leukocytes, well in line with the observed cell content seen in the IHC staining of the tissue slides.

The reproducibility experiments were performed by preparing five replicates from the same tumor. Quantitative evaluation was done by iTRAQ-LC-MS/MS. To evaluate both the variability of the preparation and the variability of the analytical LC-MS/MS workflow, one of the five preparatory replicates was divided into four equal parts and digested, labeled and quantified in parallel. The reproducibility of the workflow proved to be very good, with CVs below 15% for the entire workflow and below 10% for the LC-MS/MS part alone.

Approximately twice as many proteins were identified from the ETS samples compared with the complete lysis of the fresh frozen tissue (FF) (n=244 vs n=109). As one of the aims of the ETS method was to remove blood contamination we wanted to analyze the difference in protein content between the two sample preparations. Again, using the sequence coverage as a surrogate marker for concentration, one can note several high abundant proteins (hemoglobin B, hemoglobin Z and albumin) among the top ten proteins with the highest coverage from the FF samples, which are not present in the top ten list from the ETS.

Analyzing the GO terms related to the tissue function and structure there is a trend that proteins related to extracellular and tissue functions (GO term; cell communication, cell organization, defense response, transport, extracellular proteins and membrane proteins) are overrepresented in the FF compared with the ETS.

In addition, we observed differential quantitative expression of several previously published markers according to histological subtype (desmoplakin in squamous cell carcinoma, S100A8 and S100A9 in large cell carcinoma and galectin-3-binding protein in adenocarcinoma). As a general conclusion the ETS method proved to be efficient at removing blood contaminants, robust and well suitable for global proteomics.

2.4 GENERAL CONCLUSIONS AND FUTURE PERSPECTIVES

The overall aim of this thesis was to evaluate and optimize different stages in mass spectrometry based proteomics to facilitate biomarker discovery from clinical materials. This is obviously a very broad aim and the interventions in the individual papers have been rather different depending on the challenge at hand. The papers has included two different malignancies; AML and lung cancer, four different clinical materials; AML cells, plasma, pleural effusion and lung cancer tissue, and in addition two different mass spectrometry platforms; SELDI-TOF and MALDI-TOF/TOF. Overall this has provided an insight into the challenges and possibilities when analyzing clinical materials using global proteomics techniques. The general conclusions from this thesis could be summarized as follows;

- *Combining different analytical methods is beneficial in a proteomics workflow.*

  This is illustrated in **paper I**, where the combination of different data pre-processing methods and biomarker selection methods generated a robust set of markers. In **paper II**, by the positive effect of prefractionation in general, as well as the additional value of adding a gel-based separation step prior to the LC-MALDI-TOF/TOF more than doubled the number of identified proteins. The meta-analysis in **paper III** also proved the usefulness of combining publically available data-sets to investigate method performance. The added value of using several prefractionation methods was shown in **paper IV,** where more proteins where identified than in **paper II.** In addition, **paper IV** also benefited from the parallel analysis of FFE and IPG as well as plasma and pleural effusion for cross-validation of potential markers.

- *Using current mass spectrometry based technologies there seems to be a critical mass of identified proteins that needs to be covered to identify a large proportion of tissue leakage proteins from plasma*

  The results from the meta-analysis in **paper III** showed that the number of identified proteins rather than the separation method was the key factor to reach tissue leakage proteins in plasma. This correlated well with the results in **paper II** and **IV** where the majority of proteins are of extracellular nature.

- *Experimental design is of outmost importance in proteomics experiments when studying clinical materials.*

  Since clinical materials usually are limited both in amounts and number of patients and the sample sets contain high natural variability it is extremely important to think ahead and to do a proper optimization and experimental design before performing the experiments. This includes both data analysis and experimental factors as well as potential research questions and sample selection. Choosing the right patients, the right controls and the right clinical material for the right research question is extremely important.

Taking all results from this thesis into consideration, the most likely way forward for plasma proteomics at present would be as a validation material for candidate markers. Since plasma probably is the most common sample to have in large numbers this further highlights it's applicability for validation studies; covering a larger portion of the naturally occurring variation between individuals. Moving closer to the hypothetical marker origin, or even moving over to model systems for the biomarker discovery is probably fruitful.

However, moving away from plasma as discovery material obviously presents a drawback, if one in the end wants to measure the biomarker in plasma, as findings from model systems or tissue might not be detectable in plasma.

An obvious development that would facilitate mass spectrometry based proteomics analysis of clinical materials is technical improvement of the mass spectrometers. If accurate quantitative mass spectrometry analysis of all proteins in clinical materials were possible, much of the work in this thesis would be superfluous.

In addition to analytical depth, throughput of the analysis is crucial. To be able to cover the natural variability in/between clinical samples it is important to be able to analyze a large number of samples.

Thousands of biomarker discovery studies have been published at present, and critique that has been expressed by the scientific community is that mass spectrometry based biomarker studies results in a list of proteins, but not much more. This is in part true, validation is a large bottleneck in proteomics and validation studies on larger clinical materials, as well as molecular biology based validation studies related to function, is currently often lacking. Much information can probably be gained by data mining of currently published proteomics studies, but lack of standardized data formats and no consensus standardized public domain data repository makes this barely feasible at present [167-170].

However, as the term proteomics was coined as recently as 1995, the field is young and many quality improvements have been done over the years to raise the standards within the field, in particular on the mass spectrometry side. Examples of this include; basic quality criteria for publication of ms/ms data, instrumental improvements such as the introduction of the Orbitrap, development and improvement of data analysis tools etc. If the field continues to grow and develop as it has for the last 15 years, there is a good chance that mass spectrometry proteomics analysis of clinical materials will be done routinely within the coming 15 years.

3 ACKNOWLEDGEMENTS

Many people have helped me throughout my PhD studies. Some through hands on help in the lab, some through inspiration and motivation and some by simply making me take my mind of things for a while. This thesis would not have been possible without you.

First of all I would like to thank my main supervisor Janne Lehtiö. Your ambition and faith in what we together can accomplish has really been inspirational. You have a brilliant analytical mind and are always able to pinpoint the weaknesses in a line of reasoning. Besides that I would like to thank you for your sense of humor, your finish temper (I did deserve to be called a spoiled brat), and your courage. I would never have enjoyed my PhD studies so much without you.

Rolf Lewensohn, my co-supervisor. For always being positive and seeing the possibilities instead of the obstacles and for and making me feel welcomed in your group from day one.

Lukas Orre, my roommate and constant support. Thank you for help on everything from western blots to relationships and mortgages. I am really proud of you.

My lady-friends at KBC. Hanna and Sara. For happy and sad times, for discussions on life and everything that comes with it.

Jenny Foshed I think you make our group so much better in so many ways. You have an honest, prestigeless way of tackling things that I admire.

Luigi de Petris because you're funny, reliable and the perfect co-author.

Past and present members of the KBC crew; Kristian, Lina, Helena, Johan, Henrik, Anders and his group, Rui, Barbro, Elena, Lyris, Susanne, Per-Johan and his group, Kristina and her group, Åsa and her group and Serhiy and his group, because you make KBC such a nice place to work in.

My co-authors and collaborators; Leif, Marita, Cheun Seng, Yudi, Hirsch, Pernilla, Eva, Lena, Göran, Lotta och Per. For fruitful collaboration and for each of you learning me something new!

Birgitta Sundelin for endless walks back and forth to the thorax clinic and for preparing all the clinical samples.

Meeri help with everything from stitches, blood samples, ethical permits and administrative stuff.

Staffan for the nice cover!

Alla balla; Hannah, Karra, Helle, Emma och Chrillan. Ni är så jäkla bra. Jag är så stolt över er allihopa. Lööööööv!

Jean – för att vi är ett sånt bra team. För att du är orädd, rolig och ärlig.

Alla gamla Uppsala kompisar Ylva, Kristina, Karin, Åsa och Lina. För att jag alltid blir så glad av att träffa er!

Sara för att du förstår. Och för att du rolig, smart och omtänksam.

Emilie för att din oändliga optimism, envishet och klokhet är en ren inspiration. Viktor för att du var mitt första kompis-barn (eller barn-kompis) och alltid kommer ha en speciell plats i mitt hjärta.

Klättervännerna – min andra familj. David, Dick, Cécile, Tobbe, Aron, Marre, Katrin, Johanna W, alla medlemmar i Climb, Belay och Spotters. För all glädje, lycka, frustration, rädsla, eufori, njutning, smärta och kärlek.

Peter och Sofia för ni är så fina tillsammans.

Mamma och Pappa för trygghet, uppmuntran och förmaningar.

Stefan, du är bäst. Jag älskar dig.

4 REFERENCES

[1] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., *et al.*, Initial sequencing and analysis of the human genome. *Nature* 2001, *409*, 860-921.
[2] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., *et al.*, The sequence of the human genome. *Science* 2001, *291*, 1304-1351.
[3] Finishing the euchromatic sequence of the human genome. *Nature* 2004, *431*, 931-945.
[4] Pruitt, K. D., Harrow, J., Harte, R. A., Wallin, C., *et al.*, The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* 2009, *19*, 1316-1323.
[5] Wasinger, V. C., Cordwell, S. J., Cerpa-Poljak, A., Yan, J. X., *et al.*, Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. *Electrophoresis* 1995, *16*, 1090-1094.
[6] Gstaiger, M., Aebersold, R., Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nat Rev Genet* 2009, *10*, 617-627.
[7] Nielsen, M. L., Savitski, M. M., Zubarev, R. A., Extent of modifications in human proteome samples and their effect on dynamic range of analysis in shotgun proteomics. *Mol Cell Proteomics* 2006, *5*, 2384-2391.
[8] Anderson, N. L., Anderson, N. G., The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics* 2002, *1*, 845-867.
[9] de Godoy, L. M., Olsen, J. V., Cox, J., Nielsen, M. L., *et al.*, Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* 2008, *455*, 1251-1254.
[10] Anderson, N. L., Anderson, N. G., Pearson, T. W., Borchers, C. H., *et al.*, A human proteome detection and quantitation project. *Mol Cell Proteomics* 2009, *8*, 883-886.
[11] Minden, J. S., Dowd, S. R., Meyer, H. E., Stuhler, K., Difference gel electrophoresis. *Electrophoresis* 2009, *30 Suppl 1*, S156-161.
[12] Gorg, A., Drews, O., Luck, C., Weiland, F., Weiss, W., 2-DE with IPGs. *Electrophoresis* 2009, *30 Suppl 1*, S122-132.
[13] Iwadate, Y., Clinical proteomics in cancer research-promises and limitations of current two-dimensional gel electrophoresis. *Curr Med Chem* 2008, *15*, 2393-2400.
[14] Wingren, C., James, P., Borrebaeck, C. A., Strategy for surveying the proteome using affinity proteomics and mass spectrometry. *Proteomics* 2009, *9*, 1511-1517.
[15] Borrebaeck, C. A., Wingren, C., High-throughput proteomics using antibody microarrays: an update. *Expert Rev Mol Diagn* 2007, *7*, 673-686.
[16] Spurrier, B., Honkanen, P., Holway, A., Kumamoto, K., *et al.*, Protein and lysate array technologies in cancer research. *Biotechnol Adv* 2008, *26*, 361-369.
[17] Fernandes, T. G., Diogo, M. M., Clark, D. S., Dordick, J. S., Cabral, J. M., High-throughput cellular microarray platforms: applications in drug discovery, toxicology and stem cell research. *Trends Biotechnol* 2009, *27*, 342-349.
[18] Espina, V., Wulfkuhle, J., Liotta, L. A., Application of laser microdissection and reverse-phase protein microarrays to the molecular profiling of cancer signal pathway networks in the tissue microenvironment. *Clin Lab Med* 2009, *29*, 1-13.
[19] Camp, R. L., Neumeister, V., Rimm, D. L., A decade of tissue microarrays: progress in the discovery and validation of cancer biomarkers. *J Clin Oncol* 2008, *26*, 5630-5637.
[20] Figeys, D., Mapping the human protein interactome. *Cell Res* 2008, *18*, 716-724.
[21] Whitehouse, C. M., Dreyer, R. N., Yamashita, M., Fenn, J. B., Electrospray interface for liquid chromatographs and mass spectrometers. *Anal Chem* 1985, *57*, 675-679.
[22] Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., Whitehouse, C. M., Electrospray ionization for mass spectrometry of large biomolecules. *Science* 1989, *246*, 64-71.
[23] Karas, M., Hillenkamp, F., Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem* 1988, *60*, 2299-2301.

[24] Tanaka, K., Waki, H., Ido, Y., Akita, S., *et al.*, Protein and polymer analyses up to <I>m/z</I> 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry* 1988, *2*, 151-153.
[25] Paul, W., Steinwedel, H.S., Ein neues massenspektrometer ohne magnet feld. *Z. Naturf. A* 1954, *8*, 448-450.
[26] Siuzdak, G., *The Expanding Role of Mass Spectrometry in Biotechnology*, MCC Press, San Diego 2006.
[27] Paul, W., *Agewandte Chemie - International Edition* 1990, *29*, 739.
[28] Stephens, W. E., Proceedings of the American Physical Society. *Physical Review* 1946, *69*, 674.
[29] Comisarow, M. B., Marshall, A. G., Fourier transform ion cyclotron resonance spectroscopy. *Chem. Phys. Lett.* 1974, *25*, 282-283.
[30] Makarov, A., Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal Chem* 2000, *72*, 1156-1162.
[31] Makarov, A., Denisov, E., Lange, O., Horning, S., Dynamic range of mass accuracy in LTQ Orbitrap hybrid mass spectrometer. *J Am Soc Mass Spectrom* 2006, *17*, 977-982.
[32] Makarov, A., Denisov, E., Kholomeev, A., Balschun, W., *et al.*, Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal Chem* 2006, *78*, 2113-2120.
[33] Olsen, J. V., de Godoy, L. M., Li, G., Macek, B., *et al.*, Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol Cell Proteomics* 2005, *4*, 2010-2021.
[34] Perry, R. H., Cooks, R. G., Noll, R. J., Orbitrap mass spectrometry: instrumentation, ion motion and applications. *Mass Spectrom Rev* 2008, *27*, 661-699.
[35] Hu, Q., Noll, R. J., Li, H., Makarov, A., *et al.*, The Orbitrap: a new mass spectrometer. *J Mass Spectrom* 2005, *40*, 430-443.
[36] McLafferty, F. W., Breuker, K., Jin, M., Han, X., *et al.*, Top-down MS, a powerful complement to the high capabilities of proteolysis proteomics. *FEBS J* 2007, *274*, 6256-6268.
[37] Syka, J. E., Coon, J. J., Schroeder, M. J., Shabanowitz, J., Hunt, D. F., Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A* 2004, *101*, 9528-9533.
[38] Yildiz, P. B., Shyr, Y., Rahman, J. S., Wardwell, N. R., *et al.*, Diagnostic accuracy of MALDI mass spectrometric analysis of unfractionated serum in lung cancer. *J Thorac Oncol* 2007, *2*, 893-901.
[39] Caprioli, R. M., Farmer, T. B., Gile, J., Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. *Anal Chem* 1997, *69*, 4751-4760.
[40] Chaurand, P., Norris, J. L., Cornett, D. S., Mobley, J. A., Caprioli, R. M., New developments in profiling and imaging of proteins from tissue sections by MALDI mass spectrometry. *J Proteome Res* 2006, *5*, 2889-2900.
[41] Cornett, D. S., Mobley, J. A., Dias, E. C., Andersson, M., *et al.*, A novel histology-directed strategy for MALDI-MS tissue profiling that improves throughput and cellular specificity in human breast cancer. *Mol Cell Proteomics* 2006, *5*, 1975-1983.
[42] Chaurand, P., Stoeckli, M., Caprioli, R. M., Direct profiling of proteins in biological tissue sections by MALDI mass spectrometry. *Anal Chem* 1999, *71*, 5263-5270.
[43] Hutchens, T. W., Yip, T.-T., New desorption strategies for the mass spectrometric analysis of macromolecules. *Rapid Communications in Mass Spectrometry* 1993, *7*, 576-580.
[44] Domon, B., Aebersold, R., Mass spectrometry and protein analysis. *Science* 2006, *312*, 212-217.
[45] Jin, W. H., Dai, J., Li, S. J., Xia, Q. C., *et al.*, Human plasma proteome analysis by multidimensional chromatography prefractionation and linear ion trap mass spectrometry identification. *J Proteome Res* 2005, *4*, 613-619.

[46] Wang, H., Clouthier, S. G., Galchev, V., Misek, D. E., *et al.*, Intact-protein-based high-resolution three-dimensional quantitative analysis system for proteome profiling of biological fluids. *Mol Cell Proteomics* 2005, *4*, 618-625.
[47] Tirumalai, R. S., Chan, K. C., Prieto, D. A., Issaq, H. J., *et al.*, Characterization of the low molecular weight human serum proteome. *Mol Cell Proteomics* 2003, *2*, 1096-1103.
[48] Yang, Z., Hancock, W. S., Approach to the comprehensive analysis of glycoproteins isolated from human serum using a multi-lectin affinity column. *J Chromatogr A* 2004, *1053*, 79-88.
[49] Zhang, H., Li, X. J., Martin, D. B., Aebersold, R., Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nat Biotechnol* 2003, *21*, 660-666.
[50] Wepf, A., Glatter, T., Schmidt, A., Aebersold, R., Gstaiger, M., Quantitative interaction proteomics using mass spectrometry. *Nat Methods* 2009, *6*, 203-205.
[51] Yang, W., Steen, H., Freeman, M. R., Proteomic approaches to the analysis of multiprotein signaling complexes. *Proteomics* 2008, *8*, 832-851.
[52] Bjorhall, K., Miliotis, T., Davidsson, P., Comparison of different depletion strategies for improved resolution in proteomic analysis of human serum samples. *Proteomics* 2005, *5*, 307-317.
[53] Whiteaker, J. R., Zhang, H., Eng, J. K., Fang, R., *et al.*, Head-to-head comparison of serum fractionation techniques. *J Proteome Res* 2007, *6*, 828-836.
[54] Michel, P. E., Reymond, F., Arnaud, I. L., Josserand, J., *et al.*, Protein fractionation in a multicompartment device using Off-Gel isoelectric focusing. *Electrophoresis* 2003, *24*, 3-11.
[55] Nissum, M., Kuhfuss, S., Hauptmann, M., Obermaier, C., *et al.*, Two-dimensional separation of human plasma proteins using iterative free-flow electrophoresis. *Proteomics* 2007, *7*, 4218-4227.
[56] Wall, D. B., Kachman, M. T., Gong, S., Hinderer, R., *et al.*, Isoelectric focusing nonporous RP HPLC: a two-dimensional liquid-phase separation method for mapping of cellular proteins with identification using MALDI-TOF mass spectrometry. *Anal Chem* 2000, *72*, 1099-1111.
[57] Yang, Y., Zhang, S., Howe, K., Wilson, D. B., *et al.*, A comparison of nLC-ESI-MS/MS and nLC-MALDI-MS/MS for GeLC-based protein identification and iTRAQ-based shotgun quantitative proteomics. *J Biomol Tech* 2007, *18*, 226-237.
[58] Wolters, D. A., Washburn, M. P., Yates, J. R., 3rd, An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem* 2001, *73*, 5683-5690.
[59] Washburn, M. P., Wolters, D., Yates, J. R., 3rd, Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* 2001, *19*, 242-247.
[60] Jensen, S. S., Larsen, M. R., Evaluation of the impact of some experimental procedures on different phosphopeptide enrichment techniques. *Rapid Commun Mass Spectrom* 2007, *21*, 3635-3645.
[61] Motoyama, A., Xu, T., Ruse, C. I., Wohlschlegel, J. A., Yates, J. R., 3rd, Anion and cation mixed-bed ion exchange for enhanced multidimensional separations of peptides and phosphopeptides. *Anal Chem* 2007, *79*, 3623-3634.
[62] Hubbard, M. J., Cohen, P., On target with a new mechanism for the regulation of protein phosphorylation. *Trends Biochem Sci* 1993, *18*, 172-177.
[63] Pinkse, M. W., Uitto, P. M., Hilhorst, M. J., Ooms, B., Heck, A. J., Selective isolation at the femtomole level of phosphopeptides from proteolytic digests using 2D-NanoLC-ESI-MS/MS and titanium oxide precolumns. *Anal Chem* 2004, *76*, 3935-3943.
[64] Zhang, H., Zha, X., Tan, Y., Hornbeck, P. V., *et al.*, Phosphoprotein analysis using antibodies broadly reactive against phosphorylated motifs. *J Biol Chem* 2002, *277*, 39379-39387.
[65] Mandell, J. W., Phosphorylation state-specific antibodies: applications in investigative and diagnostic pathology. *Am J Pathol* 2003, *163*, 1687-1698.

[66] Ross, A. H., Baltimore, D., Eisen, H. N., Phosphotyrosine-containing proteins isolated by affinity chromatography with antibodies to a synthetic hapten. *Nature* 1981, *294*, 654-656.
[67] Geng, M., Ji, J., Regnier, F. E., Signature-peptide approach to detecting proteins in complex mixtures. *J Chromatogr A* 2000, *870*, 295-313.
[68] Joos, T., Affinity-MS: methods and applications in proteomics research. *Proteomics* 2009, *9*, 1418-1419.
[69] Poetz, O., Hoeppe, S., Templin, M. F., Stoll, D., Joos, T. O., Proteome wide screening using peptide affinity capture. *Proteomics* 2009, *9*, 1518-1523.
[70] Cargile, B. J., Bundy, J. L., Freeman, T. W., Stephenson, J. L., Jr., Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification. *J Proteome Res* 2004, *3*, 112-119.
[71] Cargile, B. J., Sevinsky, J. R., Essader, A. S., Stephenson, J. L., Jr., Bundy, J. L., Immobilized pH gradient isoelectric focusing as a first-dimension separation in shotgun proteomics. *J Biomol Tech* 2005, *16*, 181-189.
[72] Cargile, B. J., Talley, D. L., Stephenson, J. L., Jr., Immobilized pH gradients as a first dimension in shotgun proteomics and analysis of the accuracy of pI predictability of peptides. *Electrophoresis* 2004, *25*, 936-945.
[73] Eriksson, H., Lengqvist, J., Hedlund, J., Uhlen, K., *et al.*, Quantitative membrane proteomics applying narrow range peptide isoelectric focusing for studies of small cell lung cancer resistance mechanisms. *Proteomics* 2008, *8*, 3008-3018.
[74] Essader, A. S., Cargile, B. J., Bundy, J. L., Stephenson, J. L., Jr., A comparison of immobilized pH gradient isoelectric focusing and strong-cation-exchange chromatography as a first dimension in shotgun proteomics. *Proteomics* 2005, *5*, 24-34.
[75] Lengqvist, J., Uhlen, K., Lehtio, J., iTRAQ compatibility of peptide immobilized pH gradient isoelectric focusing. *Proteomics* 2007, *7*, 1746-1752.
[76] Heller, M., Michel, P. E., Morier, P., Crettaz, D., *et al.*, Two-stage Off-Gel isoelectric focusing: protein followed by peptide fractionation and application to proteome analysis of human plasma. *Electrophoresis* 2005, *26*, 1174-1188.
[77] Kumar, C., Mann, M., Bioinformatics analysis of mass spectrometry-based proteomics data sets. *FEBS Lett* 2009, *583*, 1703-1712.
[78] Deutsch, E. W., Lam, H., Aebersold, R., Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiol Genomics* 2008, *33*, 18-25.
[79] Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., Kuster, B., Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem* 2007, *389*, 1017-1031.
[80] Pang, J. X., Ginanni, N., Dongre, A. R., Hefta, S. A., Opitek, G. J., Biomarker discovery in urine by proteomics. *J Proteome Res* 2002, *1*, 161-169.
[81] Liu, H., Sadygov, R. G., Yates, J. R., 3rd, A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* 2004, *76*, 4193-4201.
[82] Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., *et al.*, Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* 2002, *1*, 376-386.
[83] Mann, M., Functional and quantitative proteomics using SILAC. *Nat Rev Mol Cell Biol* 2006, *7*, 952-958.
[84] Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., *et al.*, Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* 1999, *17*, 994-999.
[85] Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., *et al.*, Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* 2004, *3*, 1154-1169.
[86] Thompson, A., Schafer, J., Kuhn, K., Kienle, S., *et al.*, Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* 2003, *75*, 1895-1904.
[87] Schmidt, A., Kellermann, J., Lottspeich, F., A novel strategy for quantitative proteomics using isotope-coded protein labels. *Proteomics* 2005, *5*, 4-15.

[88] Matsuo, E., Watanabe, M., Kuyama, H., Nishimura, O., A new strategy for protein biomarker discovery utilizing 2-nitrobenzenesulfenyl (NBS) reagent and its applications to clinical samples. *J Chromatogr B Analyt Technol Biomed Life Sci* 2009, *877*, 2607-2614.
[89] Yao, X., Freas, A., Ramirez, J., Demirev, P. A., Fenselau, C., Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal Chem* 2001, *73*, 2836-2842.
[90] Reynolds, K. J., Yao, X., Fenselau, C., Proteolytic 18O labeling for comparative proteomics: evaluation of endoprotease Glu-C as the catalytic agent. *J Proteome Res* 2002, *1*, 27-33.
[91] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000, *25*, 25-29.
[92] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M., The KEGG resource for deciphering the genome. *Nucleic Acids Res* 2004, *32*, D277-280.
[93] Zhu, X., Gerstein, M., Snyder, M., Getting connected: analysis and principles of biological networks. *Genes Dev* 2007, *21*, 1010-1024.
[94] Barabasi, A. L., Oltvai, Z. N., Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004, *5*, 101-113.
[95] Ganter, B., Zidek, N., Hewitt, P. R., Muller, D., Vladimirova, A., Pathway analysis tools and toxicogenomics reference databases for risk assessment. *Pharmacogenomics* 2008, *9*, 35-54.
[96] Huang da, W., Sherman, B. T., Tan, Q., Collins, J. R., *et al.*, The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol* 2007, *8*, R183.
[97] Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., *et al.*, PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* 2003, *13*, 2129-2141.
[98] Zubarev, R. A., Nielsen, M. L., Fung, E. M., Savitski, M. M., *et al.*, Identification of dominant signaling pathways from proteomics expression data. *J Proteomics* 2008, *71*, 89-96.
[99] Alexeyenko, A., Sonnhammer, E. L., Global networks of functional coupling in eukaryotes from comprehensive data integration. *Genome Res* 2009, *19*, 1107-1116.
[100] Whiteaker, J. R., Zhang, H., Zhao, L., Wang, P., *et al.*, Integrated pipeline for mass spectrometry-based discovery and confirmation of biomarkers demonstrated in a mouse model of breast cancer. *J Proteome Res* 2007, *6*, 3962-3975.
[101] Whiteaker, J. R., Zhao, L., Zhang, H. Y., Feng, L. C., *et al.*, Antibody-based enrichment of peptides on magnetic beads for mass-spectrometry-based quantification of serum biomarkers. *Anal Biochem* 2007, *362*, 44-54.
[102] Anderson, N. L., Anderson, N. G., Haines, L. R., Hardie, D. B., *et al.*, Mass spectrometric quantitation of peptides and proteins using Stable Isotope Standards and Capture by Anti-Peptide Antibodies (SISCAPA). *J Proteome Res* 2004, *3*, 235-244.
[103] Anderson, L., Hunter, C. L., Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Mol Cell Proteomics* 2006, *5*, 573-588.
[104] Kononen, J., Bubendorf, L., Kallioniemi, A., Barlund, M., *et al.*, Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med* 1998, *4*, 844-847.
[105] Paweletz, C. P., Charboneau, L., Bichsel, V. E., Simone, N. L., *et al.*, Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene* 2001, *20*, 1981-1989.
[106] Spurrier, B., Ramalingam, S., Nishizuka, S., Reverse-phase protein lysate microarrays for cell signaling analysis. *Nat Protoc* 2008, *3*, 1796-1808.
[107] Haab, B. B., Advances in protein microarray technology for protein expression and interaction profiling. *Curr Opin Drug Discov Devel* 2001, *4*, 116-123.
[108] Cahill, D. J., Protein and antibody arrays and their medical applications. *J Immunol Methods* 2001, *250*, 81-91.
[109] Haab, B. B., Methods and applications of antibody microarrays in cancer research. *Proteomics* 2003, *3*, 2116-2122.

[110] Hanahan, D., Weinberg, R. A., The hallmarks of cancer. *Cell* 2000, *100*, 57-70.
[111] Foulds, L., The experimental study of tumor progression: a review. *Cancer Res* 1954, *14*, 327-339.
[112] Gurwitz, D., Weizman, A., Rehavi, M., Education: Teaching pharmacogenomics to prepare future physicians and researchers for personalized medicine. *Trends Pharmacol Sci* 2003, *24*, 122-125.
[113] Ludwig, J. A., Weinstein, J. N., Biomarkers in cancer staging, prognosis and treatment selection. *Nat Rev Cancer* 2005, *5*, 845-856.
[114] Andriole, G. L., Crawford, E. D., Grubb, R. L., 3rd, Buys, S. S., *et al.*, Mortality results from a randomized prostate-cancer screening trial. *N Engl J Med* 2009, *360*, 1310-1319.
[115] Stenman, U. H., Prostate-specific antigen, clinical use and staging: an overview. *Br J Urol* 1997, *79 Suppl 1*, 53-60.
[116] Thompson, I. M., Pauler, D. K., Goodman, P. J., Tangen, C. M., *et al.*, Prevalence of prostate cancer among men with a prostate-specific antigen level < or =4.0 ng per milliliter. *N Engl J Med* 2004, *350*, 2239-2246.
[117] Oesterling, J. E., Prostate specific antigen: a critical assessment of the most useful tumor marker for adenocarcinoma of the prostate. *J Urol* 1991, *145*, 907-923.
[118] Diamandis, E. P., Yu, H., Nonprostatic sources of prostate-specific antigen. *Urol Clin North Am* 1997, *24*, 275-282.
[119] Yu, H., Berkel, H., Prostate-specific antigen (PSA) in women. *J La State Med Soc* 1999, *151*, 209-213.
[120] Clarke, R., Ressom, H. W., Wang, A., Xuan, J., *et al.*, The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat Rev Cancer* 2008, *8*, 37-49.
[121] Wulfkuhle, J. D., Liotta, L. A., Petricoin, E. F., Proteomic applications for the early detection of cancer. *Nat Rev Cancer* 2003, *3*, 267-275.
[122] Glas, A. M., Floore, A., Delahaye, L. J., Witteveen, A. T., *et al.*, Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC Genomics* 2006, *7*, 278.
[123] van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., *et al.*, Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002, *415*, 530-536.
[124] Guilak, F., Alexopoulos, L. G., Haider, M. A., Ting-Beall, H. P., Setton, L. A., Zonal uniformity in mechanical properties of the chondrocyte pericellular matrix: micropipette aspiration of canine chondrons isolated by cartilage homogenization. *Ann Biomed Eng* 2005, *33*, 1312-1318.
[125] D'Souza, R., Brown, L. R., Newland, J. R., Levy, B. M., Lachman, L. B., Detection and characterization of interleukin-1 in human dental pulps. *Arch Oral Biol* 1989, *34*, 307-313.
[126] Canas, B., Pineiro, C., Calvo, E., Lopez-Ferrer, D., Gallardo, J. M., Trends in sample preparation for classical and second generation proteomics. *J Chromatogr A* 2007, *1153*, 235-258.
[127] Bodzon-Kulakowska, A., Bierczynska-Krzysik, A., Dylag, T., Drabik, A., *et al.*, Methods for samples preparation in proteomic research. *J Chromatogr B Analyt Technol Biomed Life Sci* 2007, *849*, 1-31.
[128] Hashemitabar, G. R., Razmi, G. R., Naghibi, A., Trials to induce protective immunity in mice and sheep by application of protoscolex and hydatid fluid antigen or whole body antigen of Echinococcus granulosus. *J Vet Med B Infect Dis Vet Public Health* 2005, *52*, 243-245.
[129] Radisky, D. C., Bissell, M. J., Cancer. Respect thy neighbor! *Science* 2004, *303*, 775-777.
[130] Gygi, S. P., Rochon, Y., Franza, B. R., Aebersold, R., Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* 1999, *19*, 1720-1730.
[131] Wright, M. E., Han, D. K., Aebersold, R., Mass spectrometry-based expression profiling of clinical prostate cancer. *Mol Cell Proteomics* 2005, *4*, 545-554.
[132] Becker, K. F., Schott, C., Hipp, S., Metzger, V., *et al.*, Quantitative protein analysis from formalin-fixed tissues: implications for translational clinical research and nanoscale molecular diagnosis. *J Pathol* 2007, *211*, 370-378.

[133] Lemaire, R., Desmons, A., Tabet, J. C., Day, R., *et al.*, Direct analysis and MALDI imaging of formalin-fixed, paraffin-embedded tissue sections. *J Proteome Res* 2007, *6*, 1295-1305.
[134] Groseclose, M. R., Massion, P. P., Chaurand, P., Caprioli, R. M., High-throughput proteomic analysis of formalin-fixed paraffin-embedded tissue microarrays using MALDI imaging mass spectrometry. *Proteomics* 2008, *8*, 3715-3724.
[135] Forshed, J., Pernemalm, M., Tan, C. S., Lindberg, M., *et al.*, Proteomic data analysis workflow for discovery of candidate biomarker peaks predictive of clinical outcome for patients with acute myeloid leukemia. *J Proteome Res* 2008, *7*, 2332-2341.
[136] Jacob, K., Sollier, C., Jabado, N., Circulating tumor cells: detection, molecular profiling and future prospects. *Expert Rev Proteomics* 2007, *4*, 741-756.
[137] Magi, B., Bargagli, E., Bini, L., Rottoli, P., Proteome analysis of bronchoalveolar lavage in lung diseases. *Proteomics* 2006, *6*, 6354-6369.
[138] Rapkiewicz, A., Espina, V., Zujewski, J. A., Lebowitz, P. F., *et al.*, The needle in the haystack: application of breast fine-needle aspirate samples to quantitative protein microarray technology. *Cancer* 2007, *111*, 173-184.
[139] Mustafa, D., Kros, J. M., Luider, T., Combining laser capture microdissection and proteomics techniques. *Methods Mol Biol* 2008, *428*, 159-178.
[140] Omenn, G. S., The Human Proteome Organization Plasma Proteome Project pilot phase: reference specimens, technology platform comparisons, and standardized data submissions and analyses. *Proteomics* 2004, *4*, 1235-1240.
[141] Omenn, G. S., International collaboration in clinical chemistry and laboratory medicine: the Human Proteome Organization (HUPO) Plasma Proteome Project. *Clin Chem Lab Med* 2004, *42*, 1-2.
[142] States, D. J., Omenn, G. S., Blackwell, T. W., Fermin, D., *et al.*, Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. *Nat Biotechnol* 2006, *24*, 333-338.
[143] Schiess, R., Wollscheid, B., Aebersold, R., Targeted proteomic strategy for clinical biomarker discovery. *Mol Oncol* 2009, *3*, 33-44.
[144] Rifai, N., Gillette, M. A., Carr, S. A., Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat Biotechnol* 2006, *24*, 971-983.
[145] Westman-Brinkmalm, A., Ruetschi, U., Portelius, E., Andreasson, U., *et al.*, Proteomics/peptidomics tools to find CSF biomarkers for neurodegenerative diseases. *Front Biosci* 2009, *14*, 1793-1806.
[146] Pan, S., Zhu, D., Quinn, J. F., Peskind, E. R., *et al.*, A combined dataset of human cerebrospinal fluid proteins identified by multi-dimensional chromatography and tandem mass spectrometry. *Proteomics* 2007, *7*, 469-473.
[147] Farina, A., Dumonceau, J. M., Lescuyer, P., Proteomic analysis of human bile and potential applications for cancer diagnosis. *Expert Rev Proteomics* 2009, *6*, 285-301.
[148] Farina, A., Dumonceau, J. M., Frossard, J. L., Hadengue, A., *et al.*, Proteomic analysis of human bile from malignant biliary stenosis induced by pancreatic cancer. *J Proteome Res* 2009, *8*, 159-169.
[149] Choolani, M., Narasimhan, K., Kolla, V., Hahn, S., Proteomic technologies for prenatal diagnostics: advances and challenges ahead. *Expert Rev Proteomics* 2009, *6*, 87-101.
[150] Buhimschi, I. A., Zhao, G., Rosenberg, V. A., Abdel-Razeq, S., *et al.*, Multidimensional proteomics analysis of amniotic fluid to provide insight into the mechanisms of idiopathic preterm birth. *PLoS One* 2008, *3*, e2049.
[151] Hu, S., Loo, J. A., Wong, D. T., Human saliva proteome analysis and disease biomarker discovery. *Expert Rev Proteomics* 2007, *4*, 531-538.
[152] Rao, P. V., Reddy, A. P., Lu, X., Dasari, S., *et al.*, Proteomic identification of salivary biomarkers of type-2 diabetes. *J Proteome Res* 2009, *8*, 239-245.
[153] Wilson, R., Whitelock, J. M., Bateman, J. F., Proteomics makes progress in cartilage and arthritis research. *Matrix Biol* 2009, *28*, 121-128.
[154] Chang, X., Cui, Y., Zong, M., Zhao, Y., *et al.*, Identification of proteins with increased expression in rheumatoid arthritis synovial tissues. *J Rheumatol* 2009, *36*, 872-880.

[155] Hu, S., Loo, J. A., Wong, D. T., Human body fluid proteome analysis. *Proteomics* 2006, *6*, 6326-6353.
[156] Zhou, L., Beuerman, R. W., Chan, C. M., Zhao, S. Z., *et al.*, Identification of Tear Fluid Biomarkers in Dry Eye Syndrome Using iTRAQ Quantitative Proteomics. *J Proteome Res* 2009.
[157] Mannello, F., Medda, V., Tonti, G. A., Protein profile analysis of the breast microenvironment to differentiate healthy women from breast cancer patients. *Expert Rev Proteomics* 2009, *6*, 43-60.
[158] Pawlik, T. M., Hawke, D. H., Liu, Y., Krishnamurthy, S., *et al.*, Proteomic analysis of nipple aspirate fluid from women with early-stage breast cancer using isotope-coded affinity tags and tandem mass spectrometry reveals differential expression of vitamin D binding protein. *BMC Cancer* 2006, *6*, 68.
[159] Soltermann, A., Ossola, R., Kilgus-Hawelski, S., von Eckardstein, A., *et al.*, N-glycoprotein profiling of lung adenocarcinoma pleural effusions by shotgun proteomics. *Cancer* 2008, *114*, 124-133.
[160] Pernemalm, M., De Petris, L., Eriksson, H., Branden, E., *et al.*, Use of narrow-range peptide IEF to improve detection of lung adenocarcinoma markers in plasma and pleural effusion. *Proteomics* 2009, *9*, 3414-3424.
[161] Omenn, G. S., States, D. J., Adamski, M., Blackwell, T. W., *et al.*, Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* 2005, *5*, 3226-3245.
[162] Ow, S. Y., Salim, M., Noirel, J., Evans, C., *et al.*, iTRAQ underestimation in simple and complex mixtures: The good, the bad and the ugly. *J Proteome Res* 2009.
[163] De Petris, L., Orre, L. M., Kanter, L., Pernemalm, M., *et al.*, Tumor expression of S100A6 correlates with survival of patients with stage I non-small-cell lung cancer. *Lung Cancer* 2009, *63*, 410-417.
[164] Orre, L. M., Pernemalm, M., Lengqvist, J., Lewensohn, R., Lehtio, J., Up-regulation, modification, and translocation of S100A6 induced by exposure to ionizing radiation revealed by proteomics profiling. *Mol Cell Proteomics* 2007, *6*, 2122-2131.
[165] Tan, C. S., Ploner, A., Quandt, A., Lehtio, J., *et al.*, Annotated regions of significance of SELDI-TOF-MS spectra for detecting protein biomarkers. *Proteomics* 2006, *6*, 6124-6133.
[166] Tan, C. S., Ploner, A., Quandt, A., Lehtio, J., Pawitan, Y., Finding regions of significance in SELDI measurements for identifying protein biomarkers. *Bioinformatics* 2006, *22*, 1515-1523.
[167] Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., *et al.*, The PeptideAtlas project. *Nucleic Acids Res* 2006, *34*, D655-658.
[168] Jones, P., Cote, R., The PRIDE proteomics identifications database: data submission, query, and dataset comparison. *Methods Mol Biol* 2008, *484*, 287-303.
[169] Jones, P., Cote, R. G., Cho, S. Y., Klie, S., *et al.*, PRIDE: new developments and new datasets. *Nucleic Acids Res* 2008, *36*, D878-883.
[170] Prince, J. T., Carlson, M. W., Wang, R., Lu, P., Marcotte, E. M., The need for a public proteomics repository. *Nat Biotechnol* 2004, *22*, 471-472.