From DEPARTMENT OF ONCOLOGY-PATHOLOGY
SCIENCE FOR LIFE LABORATORY
Karolinska Institutet, Stockholm, Sweden

MULTIVARIATE ANALYSIS OF CANCER PROTEOMICS DATA – TOWARDS A BIOLOGICAL SYSTEMS VIEW AND UNDERSTANDING

Lina Hultin Rosenberg

Stockholm 2013

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

Cover artwork by Jakob Hultin Rosenberg.

Published by Karolinska Institutet. Printed by Larserics Digital Print AB.

ISBN 978-91-7549-243-8

ABSTRACT

Important aims of cancer proteomics include gaining better understanding of cancer biology and identifying cancer biomarkers. Mass spectrometry (MS) based shotgun proteomics allow for identification and quantification of thousands of proteins in complex human samples. However, proteomics discovery research in clinical material faces many challenges. The biological differences between groups are often expected to be rather small, at the same time the human proteome is highly complex and there is large biological variation between clinical samples. To be able to extract meaningful results from proteomics data derived from biological and clinical material, care has to be taken to all the critical steps in the data analysis workflow. First of all we need to have robust methods to extract good quality data. A proper statistical analysis is then of outmost importance, taking into account risks of over-fitting and false positives. In addition, we also need system based approaches to relate the data to clinical and biological questions.

The main goal of this thesis was to generate robust methods for selection of key proteins, networks and pathways relevant for answering biological and clinical questions. The work includes development and evaluation of workflows for quantitative analysis of proteomics data.

In **paper I**, a multivariate meta-analysis workflow was developed to link existing proteomics data from human colon and prostate tumours. The aim was to identify proteins distinguishing between normal and tumour samples independent of tissue origin, as well as to find unique markers. The bioinformatics workflow for meta-analysis developed in this study enabled the finding of a common protein profile for the two malign tumour types, which was not possible when analysing the data sets separately. The purpose of **paper II** was to generate a basis for the decision of what protein quantities are reliable and find a way for accurate and precise protein quantification. We developed a methodology for improved protein quantification in shotgun proteomics and introduced a way to assess quantification for proteins with few peptides. The experimental design and developed algorithms decreased the relative protein quantification error in the analysis of complex biological samples. In **paper III**, we presented SpliceVista, a tool for splice variant identification and visualization based on MS proteomics data. SpliceVista identifies splice variant specific peptides and provides the possibility to perform splice variant specific quantitative analysis. SpliceVista was applied in two experimental datasets to exemplify its capability of detecting differentially expressed splice variants at the protein level. The aim of **paper IV** was to develop a network based analysis workflow for proteomics data to identify protein subnetworks with different activity between groups of samples. The methodology, which is based on a multivariate model directed by the network, was applied to several of our clinical mass spectrometry datasets. The output from the subnetwork analysis was functional subunits of proteins, rather than a collection of sparse proteins, which were shown to more readily provide a model of the biological mechanisms studied, and thus aid in the biological interpretation.

LIST OF PUBLICATIONS

I. **Rosenberg LH**, Franzén B, Auer G, Lehtiö J, Forshed J. Multivariate meta-analysis of proteomics data from human prostate and colon tumours. *BMC Bioinformatics* 2010 Sep 17;11:468

Contribution: I was taking part in the planning and the design of the study. I was responsible for the design of the workflow and relevant methods as well as for the development and implementation of the statistical analysis. I performed the analysis of the data and the validation of the statistical models. I compiled the results, generated figures and wrote the manuscript with support from the other authors.

II. **Hultin-Rosenberg L**, Forshed J, Branca R M M, Lehtiö J and Johansson H. Defining, comparing and improving iTRAQ quantification in mass spectrometry proteomics data. *Molecular and Cellular Proteomics* 2013, 12.7

Contribution: I planned, together with author 2, the statistical analysis of the data. I developed and implemented the methods for analysis of the data as well as generated the results and figures. I compiled the results and figures and wrote the manuscript.

III. Zhu Y, **Hultin-Rosenberg L**, Forshed J, Lehtiö J. SpliceVista - an identification and visualization tool to detect splice variants in shotgun proteomics data. *Manuscript submitted to Molecular and Cellular Proteomics*

Contribution: All authors contributed with idea and design of the study. I participated in the planning of the study. The first author developed the tools and methods for the analysis, with support and input from me. I supported in compiling the results and figures and took a big part in the manuscript writing.

IV. **Hultin-Rosenberg L**, Zhu Y, Branca R M M, Eriksson H, Forshed J, Lehtiö J. A multivariate network based analysis of in-depth proteomics data from cancer studies. *Manuscript*

Contribution: I came up with the idea of the study together with authors 5 and 6. I designed and planned the study. I developed and implemented the methods as well as the evaluation of the results, with help from author 5. I performed the analysis and generated figures. I wrote the manuscript.

Publication not included in this thesis

Sofiadis A, Becker S, Hellman U, **Hultin-Rosenberg L**, Dinets A, Hulchiy M, Zedenius J, Wallin G, Foukakis T, Höög A, Auer G, Lehtiö J, Larsson C. Proteomic profiling of follicular and papillary thyroid tumors. *Eur J Endocrinol.* 2012 Apr;166(4):657-67.

CONTENTS

LIST OF ABBREVIATIONS

2DE	Two dimensional gel electrophoresis
ANOVA	Analysis of variance
CID	Collision induced dissociation
ESI	Electrospray ionization
EVDB	Evidence viewer database
FDR	False discovery rate
FWER	Family wise error rate
HCD	Higher-energy collisional dissociation
ICAT	Isotope coded affinity tag
ICR	Ion cyclotron resonance
IPG-IEF	Immobilized pH gradient isoelectric focusing
IT	Ion trap
iTRAQ	Isobaric tags for relative and absolute quantification
LC	Liquid chromatography
LOO	Leave one out
LTQ	Linear quadrupole ion trap
MALDI	Matrix assisted laser desorption ionization
MS	Mass spectrometry
MSigDB	Molecular signatures database
Mw	Molecular weight
OPLS	Orthogonal projection to latent structures
PCA	Principal component analysis
pI	Isoelectric point
PLS	Partial least squares or Projection to latent structures
PLS-DA	Partial least squares discriminant analysis
PQPQ	Protein quantification by peptide quality control
PSM	Peptide spectrum match
Q	Quadrupole
RMSE	Root mean square error
RSD	Relative standard deviation
SILAC	Stable isotope labelling by amino acids in cell culture
SRM	Selected reaction monitoring
STRING	Search tool for the retrieval of interacting genes
SVSP	Splice variant specific peptide
TMT	Tandem mass tags
TOF	Time of flight
VIP	Variable importance on projection

1 BACKGROUND

1.1 PROTEOMICS

The proteome is the entire set of proteins expressed by a genome, cell, tissue or organism at a certain time, under certain conditions [1, 2]. The term proteomics describes the large-scale study of the proteome; including protein composition, protein structure, expression, function and interactions. The Human Genome Project [3, 4] provided a blueprint for the gene-encoded proteins potentially active in human cells, but there is still limited knowledge on the majority of the around 20 000 protein-coding genes. The Human Proteome Project [5, 6] was launched in 2010 with the goal of mapping the entire human proteome. Doing this is a formidable task, the total number of different proteins in the human proteome is estimated to be around 1 million [7, 8] (Figure 1). Further, in contrast to the genome, the proteome is much more dynamic and in constant change. Proteins are expressed at distinct times, in distinct cell types and only under certain conditions, as well as undergo differential splicing and post-translational modifications. This means that even the basic set of proteins that are produced in a cell needs to be determined.

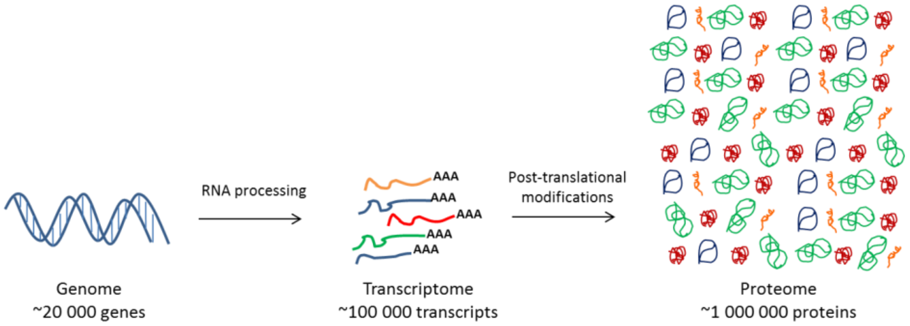

Figure 1. The DNA to RNA to protein complexity. Each gene can give rise to multiple mRNA transcripts by using alternative promoters, alternative transcription termination sites, alternative splicing and mRNA editing. The number of different protein variants from one gene is further increased by the various protein post-translational modifications.

The mRNA level is often measured as a proxy for the protein levels. Studies of differential mRNA expression are informative, but the mRNA level has been found to have limited correlation with the protein level [9, 10]. It is now known that mRNA is not always translated into protein, and the amount of protein produced for a given amount of mRNA depends both on the individual gene and on the current physiological state of the cell. Differences in protein synthesis and degradation also complicate the comparison, as mRNA and protein levels result from the coupled processes of synthesis and degradation. In addition, studies of RNA levels have limitations regarding information on protein function and interaction, and lacks information on post-translational modifications. Proteomics experiments confirm the presence of the specific protein and provide a direct measure of the protein quantity in a cell at a given time and condition. Another advantage of proteomics is that often the identified protein is the biological executive unit.

1.1.1 Cancer proteomics

Cancer proteomics is the study of protein changes related to cancer. For revealing signalling pathways causing cancer or other diseases, protein level measurements are particularly informative since protein mediated signalling controls the majority of cellular events. The importance of proteins in human diseases can further be illustrated by the fact that a majority of all drugs are targeted to have an effect on proteins [11].

Tumorigenesis in humans is a multi-step process; the steps reflect genetic alterations that drive the transformation of a normal cell into a cancer cell. The hallmarks of cancer comprise six biological capabilities, essential for the development of malignant cancer: sustaining proliferative signalling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis [12]. Two additional emerging hallmarks have been proposed: reprogramming of energy metabolism and evading immune destruction [13]. During the transformation of a normal cell into a malignant cell, several changes occur at the protein level, including altered expression, differential protein modification, as well as changes in activity and localization. Identifying and understanding these changes is the main goal in cancer proteomics [14, 15]. Despite major progresses in detection and therapy, cancer remains a major public health challenge. Cancer is a leading cause of death worldwide, about 12.7 million cancer cases and 7.6 million cancer deaths were estimated for year 2008 [16]. A better understanding of the development of drug resistance, as well as development of biomarkers for the early detection and selection of the most effective therapeutic strategies, are urgently needed [17, 18].

1.1.2 Clinical and Biological discovery research

The purposes of studying the proteome in relation to cancer can be several, but there are two main starting points: to gain better understanding of the cancer biology or to identify cancer biomarkers. Biological studies are often performed in model systems such as cell lines or animal models, while biomarker studies are preferably performed in clinical materials. Biomarkers are biological molecules that are indicators of a biological state. A biomarker can be used to provide an early indication or detection of the disease (diagnostic marker), to monitor disease progression and tell something about the disease outcome (prognostic marker) and to tell how a patient will respond to a treatment (predictive marker) [18]. Biomarkers based on the individual genetic make-up of patients can be used to design tailored treatment, an approach called personalized medicine. Personalized medicine is very important for cancer treatment, since population based medicine has not been successful for many cancer types [19]. Currently, it is very difficult to predict which patient will respond well to a treatment, as tumours often develop resistance to drugs. Development of therapy related biomarkers to select the most effective treatment, as well as diagnostic biomarkers to enable early diagnosis are key aspects to improve prognosis and survival for cancer patients [14, 15].

Recent advances in genomics and proteomics technologies have gained a lot of interest and expectations in the quest for cancer biomarkers. Unfortunately, the biomarker discovery research has so far mainly failed to deliver biomarkers for clinical use [19]. Omics technologies such as proteomics and DNA microarrays have generated more than 150 000 papers on putative biomarkers, but less than 100 have been validated for clinical practice [20].

Proteomics discovery research in clinical material faces many challenges [21, 22]. A big challenge is small sample cohorts in combination with large and unknown complexity of the human proteome as well as large biological variation between clinical samples. This is due to normal variation between healthy individuals as well as disease heterogeneity. Another complicating factor is that the differences in protein levels between groups might be very small, sometimes even smaller than the normal biological variation. Further, the low concentration of potential protein biomarkers [23, 24] makes biomarker discovery difficult, since most proteomics technologies are biased towards the detection of high abundant proteins. A number of key factors causing biomarker discovery to fail have been identified [19, 20]. Many of those can be explained by the influence of bias, the existence of a hidden structure in the data making the marker appear promising. Bias has been suggested to be the single biggest threat to validity of biomarker studies [25, 26], mainly because there are so many different sources of bias that can be difficult to keep control of. Furthermore, the observational design used in biomarker research is more subject to bias; subjects are selected and not randomly assigned, and baseline equality between cases and controls can most often not be assured. Many other factors can be explained by the use of inappropriate statistical methods. In high dimensional data there is a risk of over-fitting the statistical model to the data, giving overly optimistic results. Another risk of analysing data with thousands of variables, and often few subjects, are the false positives, positive results that occur just by random events.

Failures in biomarker development cost a lot, in terms of money, time, labour, talent, and reliability for the research field. To overcome the main obstacles for biomarker research and to increase the chances of taking a biomarker into clinic, the study has to be planned and executed carefully [21]. The selection of samples to include and a valid experimental design trying to avoid any possible bias is crucial. The experimental platform need to be suitable for the type of material and measurements, and the performance of the assay, in terms of sensitivity, accuracy and robustness, should be known. The technical, experimental and biological variation of the system has to be assessed and considered in the handling of the quantitative data. A proper statistical analysis is of outmost importance, taking into account risks of over-fitting and false positives. Testing thousands of hypotheses simultaneously requires methods for multiple testing correction to keep control of the false discovery rate. Further, the statistical model and the biomarker have to be validated properly, preferably in an independent sample cohort using an orthogonal technique. Hence the analytical properties of the validation method also have to be taken into account.

1.2 PROTEOMICS TECHNOLOGIES

As mentioned above, there are several analytical challenges in studying the human proteome, as compared to studying the genome (DNA) or the transcriptome (RNA). Besides the size and complexity of the human proteome, there are large differences in protein abundance, spanning over ten orders of magnitude in human plasma [23, 27] and at least six in tissue. Proteins are also chemically more heterogeneous than DNA and RNA and differ largely in solubility, size and p*I*. Those challenges put high demands on the methods used for proteomics analysis, to be able to cover as much as possible of the human proteome and also be able to reach the low abundant proteins.

The main analytical techniques aiming at studying the proteome have traditionally been two-dimensional gel electrophoresis (2DE) together with mass spectrometry (MS). In 2DE, proteins are separated in two dimensions based on isoelectric point (p*I*) and size (Mw) (Figure 2A) [28]. The gel is stained and protein spots of interest can be cut out and identified using MS. The 2DE technique has limitations such as limited throughput, low dynamic range and low resolution; a typical 2DE experiment detects approximately 2000-3000 protein spots out of which only a subset will be identified [29-32]. Today liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) is a commonly used method to study protein expression on a proteome/genome wide scale. The peptides, proteins or other analytes eluting from the LC column are separated according to their mass-to-charge (m/z) ratio by the mass spectrometer (Figure 2B). Recent developments in methods and instruments for mass spectrometry enable large scale quantitative proteomics analysis of complex samples with very good coverage [10, 33-42]. The number of samples feasible to analyse by MS is however limited by low throughput. At present, some publications have reported over 10 000 proteins identified and quantified in human cell lines [43-45]. The developments have also enabled the quantification of complete proteomes of model organisms such as yeast [46-49]. The technical advances have moreover made MS based proteomics an important tool for biomarker discovery [24, 50-58]. In addition to mass spectrometry methods, affinity based proteomics methods using antibodies are also widely used to study protein levels, protein localization and protein interactions [59-63].

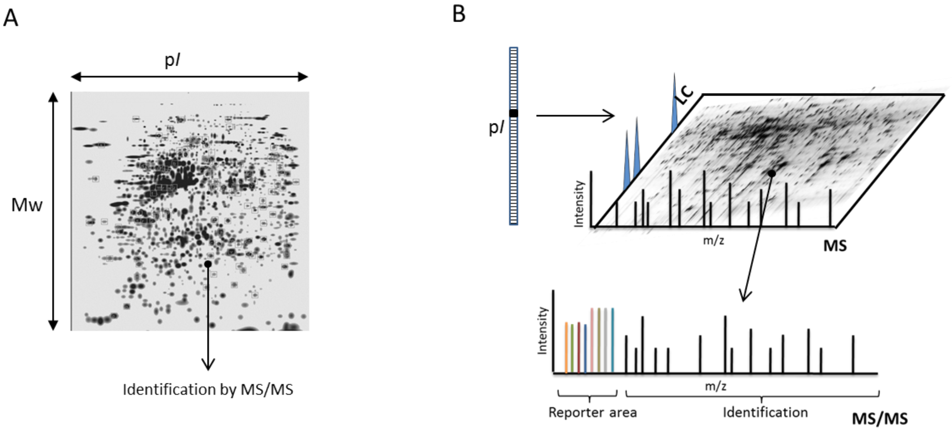

Figure 2. A: Two-dimensional gel electrophoresis. Proteins are separated in two dimensions based on isoelectric point (p*I*) and size (Mw). Protein spots of interest can be cut out and identified using MS/MS. B: Liquid chromatography coupled to tandem mass spectrometry. Peptides are usually fractionated by charge or here by p*I*, followed by separation by hydrophobicity (retention time in LC column) and by mass-to-charge ratio (m/z) in tandem mass spectrometry, first on peptide ions in MS1 and secondly on fragmented peptides in MS2 (see Figure 3).

1.3 MASS SPECTROMETRY

Mass spectrometry (MS) is an analytical technique that separates molecules according to their mass-to-charge ratio (m/z) [35, 64, 65]. It can be used for determining masses of particles, determining the elemental composition of a sample or molecule, and for elucidating the chemical structures of molecules, such as peptides, proteins, metabolites or other chemical compounds. In proteomics, MS can be used for protein quantification, protein identification, identification of protein modifications and protein complexes, as well as protein localization (imaging) [36, 66-68].

MS instruments consist of three major modules: an *ion source*, a *mass analyser* and a *detector* (Figure 3). In the *ion source*, the analytes are ionized and brought into gas phase. The most commonly used ion sources in proteomics are electrospray ionization (ESI) or matrix assisted laser desorption ionization (MALDI). The *mass analyser* separates the ions based on their m/z ratio, by applying electromagnetic fields.

Orbitrap and ion cyclotron resonance (ICR) separate ions based on m/z resonance frequency, quadrupoles (Q) and ion traps (IT) separate ions based on stability of their paths in oscillating electric fields and time of flight (TOF) analysers use flight time. Once separated by m/z, the *detector* measures the number of ions hitting the detector and provides the data for calculating the abundance of each ion cloud present. The ion signal is processed into a mass spectrum, with m/z on x-axis and ion count on y-axis.

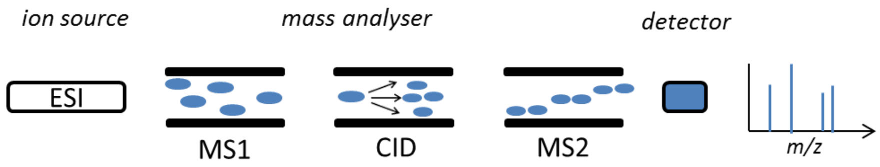

Figure 3. Major modules of tandem mass spectrometers. In the ion source the analytes are ionized. The mass analyser separates the ions based on m/z ratio. In shotgun proteomics, the first analyser (MS1) separates the peptide ions (precursor ions), peptides are then fragmented by collision energy (here exemplified by CID, collision induced dissociation) and the fragment ions (product ions) are separated by the second analyser (MS2). The detector measures the number of ions for a certain m/z ratio, which is used for the generation of the mass spectrum.

1.3.1 Shotgun proteomics

There are two main strategies for mass spectrometry based proteomics: bottom-up and top-down. In a top-down approach, intact proteins are analysed directly by mass spectrometry. Measuring intact proteins directly in MS on a larger scale is limited to rather small proteins (<45 kDa). Another problem of analysing intact proteins is that they have very different properties, making some proteins difficult to solubilize, separate and ionize by MS.

In a bottom-up approach, proteins are enzymatically digested into peptides, which are analysed by mass spectrometry [7, 40, 69-71]. Bottom-up, also known as shotgun proteomics, is the far most common workflow in MS based proteomics. A typical quantitative shotgun proteomics workflow in our lab is depicted in Figure 4.

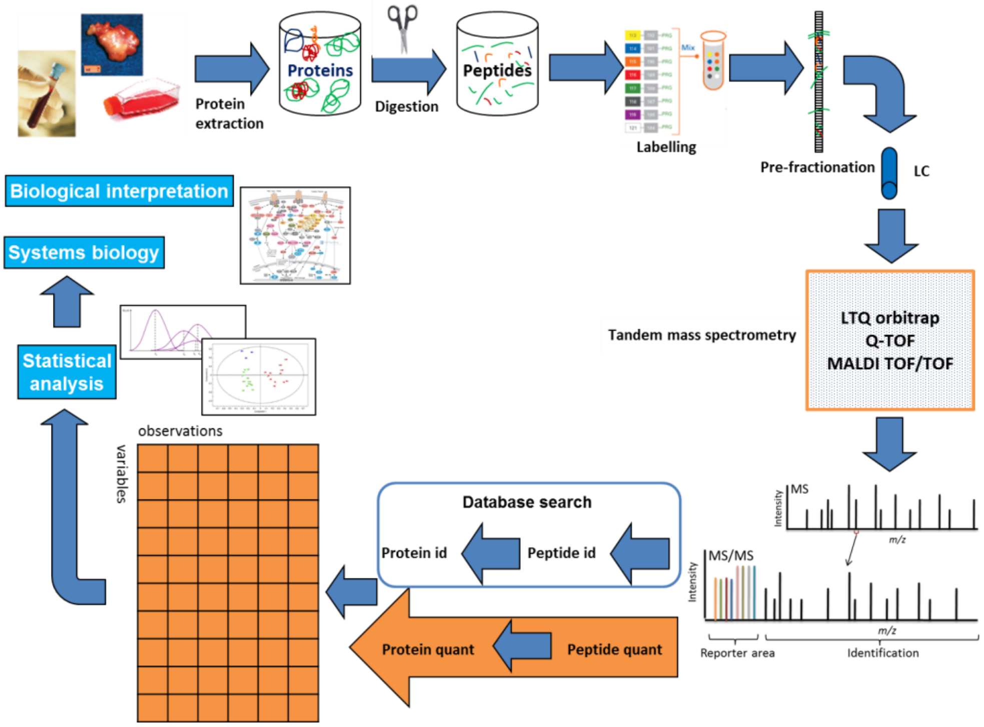

Figure 4. Quantitative shotgun proteomics workflow. Proteins are digested to produce a peptide mixture. The peptides are labelled, pooled and pre-fractionated. Fractions are loaded onto a nano column and the peptides are separated by reverse phase chromatography. As the peptides elute from the column, they are subject to tandem mass spectrometry analysis. The MS1 and MS2 spectra yield peptide identifications, which are used to infer proteins. The reporter ions from the labels are used for relative quantification of peptides. Peptide ratios are then summarized into protein ratios. Statistical analysis is performed to select the most important proteins for further systems biology based analysis to infer biological interpretation.

Shotgun proteomics is based on *enzymatic cleavage* of proteins into peptides (usually by trypsin). This is performed to facilitate ionization and fragmentation. It further avoids problems associated with intact protein analysis, such as poor separation efficiency and poor sensitivity.

The level of sample complexity, and protein abundance range, influences the performance of the MS analysis. It is difficult to obtain optimal ionization and fragmentation process for all analytes in complex samples, since the analytes have very different chemical properties. In addition, MS has limited dynamic range of detection, limiting sensitivity and quantification in complex samples. To overcome these challenges, and to maximize protein identifications, various steps directed at reducing sample complexity can be performed prior to MS analysis [64, 72, 73].

The most common approach to reduce sample complexity is by *pre-fractionation*, performed either on protein or peptide level. Since the sample complexity is increased by enzymatic cleavage (by a factor of about 40), pre-fractionation on peptide level is particularly valuable. By pre-fractionation, the peptide sample can be fractionated according to its physicochemical properties such as charge, isoelectric point, hydrophobicity or a combination of these. Alternatively, specific subsets of the sample can be targeted through enrichment of peptides or proteins using affinity-based resins or antibody-based immunoprecipitation.

To further reduce complexity of the peptide mixture, the peptides are subjected to *separation* prior to MS analysis. The separation is required to detect low-abundance proteins that would otherwise be overshadowed by higher abundance signal, as well as for un-ambiguous identifications. A common setup is to couple a liquid chromatography system (LC) to a mass spectrometer.

Peptides eluting from the LC column are then analysed by *tandem MS*. The peptides are ionized and analysed by the first MS, generating the peptide ion spectrum (MS1). For each MS1 scan, the top 5-10 peaks are usually fragmented and subjected to the second MS scan generating the fragment ion spectrum (MS2) [74].

The raw data (MS1 and MS2 spectra) generated by the instrument is first processed by signal processing softwares to reduce the raw data into a set of peaks [75, 76].

Peptide identification is typically accomplished by matching the experimental MS2 spectra to in-silico predicted spectra generated by a theoretical digest of a *protein database*, using the precursor ion mass as support. The peptide spectrum matches (PSMs) are then summarized into protein identities.

Peptide quantification can be performed either based on the MS1 spectra, as in label-free quantification, or based on the MS2 spectra, as in quantification based on isobaric labels.

The major advantage of shotgun proteomics is the ability to identify and quantify thousands of proteins in a single analysis. One disadvantage is the informatics challenges related to processing the large amount of acquired data [76]. A shotgun proteomics approach is most suitable for discovery projects aiming at rapid identification, and relative quantification, of complex sample mixtures in a limited number of samples. It is a hypothesis generating experiment that requires several follow up steps using alternative techniques. Typically, the proteins identified in the discovery phase would be validated with more targeted approaches like selected reaction monitoring (SRM) mass spectrometry [77-80] or affinity based proteomics [63, 81] in a larger cohort.

1.3.1.1 Labelling and Quantification

Many proteomics studies aim at studying differences in protein expression levels between different conditions. Such comparative analysis depends on protein quantification [69]. As mentioned before, two principally different approaches exist for quantification by MS: label free methods and methods based on stable isotope labelling [82-85].

Label free quantification is based either on the mass spectrometric signal intensity in MS1 for any given peptide, or on spectral counting; using the number of times the peptides from certain proteins are detected as a proxy for protein abundance [86-89].

An advantage of isotopic labelling is that it enables pooling of samples, thereby reducing MS run time as well as technical variability. ICAT [90], iTRAQ [91], TMT [92] and SILAC [93] are among the most commonly used labelling methods based on stable isotopes. iTRAQ (isobaric tags for relative and absolute quantification) allow for simultaneous relative quantification of up to eight samples within a single run. Using iTRAQ labelling, fragmented reporter ions from the tag are used for relative quantification in MS/MS mode.

1.3.1.2 Peptide identification

Modern MS instruments generate an enormous amount of fragment ion spectra per hour of data acquisition. The fragment ion spectrum of a peptide ion needs to be assigned to a peptide sequence. There are several different computational approaches to do this [94-98]: i) Database searching; experimental fragment ion spectra are matched with predicted spectra based on theoretical digest of protein sequences. Experimental spectra can also be used; in this case fragment ion spectra are matched to libraries containing experimental MS/MS spectra identified in previous experiments (Spectral library search), ii) De novo sequencing; peptide sequences are explicitly read out directly from the fragment ion spectra, iii) Hybrid approaches; database searching assisted by de novo sequencing.

1.3.1.3 Protein identification

The purpose of most proteomics experiments is the identification, and quantification, of the proteins present in the sample prior to digestion. Peptide sequences identified by a shotgun proteomics experiment thus need to be assembled into proteins. This is not straightforward due to several reasons [99, 100]. The major cause is peptide sequences shared between several proteins, leading to ambiguities in the identification process. The presence of several proteoforms (protein variants) [101], often with very similar protein sequences, further complicates the process. Commonly the parsimony principle (Occam’s razor) is used to infer proteins [102]; it determines the smallest number of proteins that can account for all observed peptides.

1.3.1.4 Proteoforms

Protein post-translational modification increases the functional diversity of the proteome by the covalent addition of functional groups to proteins, proteolytic cleavage of regulatory subunits or degradation of entire proteins. Post-translational modifications are key mechanisms to increase proteomic diversity. The total number of protein variants in the human proteome is estimated at over 1 million [8]. Protein isoforms also arise due to alternative splicing of the mRNA. Eukaryotic genes consist of exonic (protein coding) and intronic (non-coding) regions, after transcription the introns are removed by a process called splicing. Alternative splicing allows for the production of a variety of different proteins from one gene, by splicing and reconnecting exonic sequences in alternative ways to produce mature mRNA [103]. Alternative splicing is a very prevalent process in the human genome, it is estimated that around 92-94% of human genes has the potential to undergo alternative splicing [104]. It thus has great significance in increasing the proteome diversity and complexity. Alternative splicing plays an important role in regulating gene expression; it determines binding properties, intracellular localization, enzymatic activity, protein stability and posttranslational modifications of a large number of proteins. Disruption of alternative splicing events has been implicated in a large number of diseases, such as neurodegenerative, cardiovascular, respiratory and metabolic diseases, as well as several cancer types [105-107].

1.4 QUANTITATIVE DATA ANALYSIS

Once the peptide sequences have been determined and assembled into protein groups and the quantitative measurements have been defined both on peptide and protein level (see more details in sections 2.2.6 and 2.2.7), the analysis of the quantitative MS data can take place.

To extract meaningful results from MS proteomics data on biological and clinical material, care has to be taken to all the critical steps in the quantitative data analysis workflow, from handling the raw data to the statistical analysis and the biological interpretation of the result [70, 108-111]. The data analysis part is specifically delicate in large scale omics experiments since often thousands of variables are measured for only a few samples [112]. This implies a risk in the statistical analysis step to generate false positive discoveries. Further, the data often harbours a large amount of noise, in terms of biological variation, technical variation as well as experimental variation. The first important step is to know the quality of the quantitative data, to be aware of the limitations and reasonable expectations. Secondly, a suitable statistical method has to be selected and the statistical validation of the result has to be done carefully.

1.4.1 Pre-processing and Quality control

Prior to any statistical analysis the quantitative mass spectrometry data has to be *pre-processed* and *quality controlled*. The quantitative data from an iTRAQ experiment is expressed as ratios between iTRAQ channels, since the iTRAQ reporter ions are used for relative quantification of each peptide. Often the ratios are log transformed to give the up- and down regulations equal importance, prior to statistical analysis. Further, the quantitative data need to be normalized to make samples and experiments comparable. The amount of missing values in the data also has to be assessed and proteins with a large amount of missing data points might have to be excluded prior to statistical analysis.

Quality control of the quantitative data also has to be performed; this can be done by investigating the distribution of the data and missing values. A useful plot for investigation of data distribution, data separation and outliers is the PCA (Principal Component Analysis) scores plot (see section 1.4.2.3). For detecting differences in the distribution between samples or pools of samples, the boxplot can be very useful (see example in Figure 5). The purpose of the quality control is to detect problems in the quantitative data such as biases in data distribution, which can then be adjusted for by normalization prior to statistical analysis.

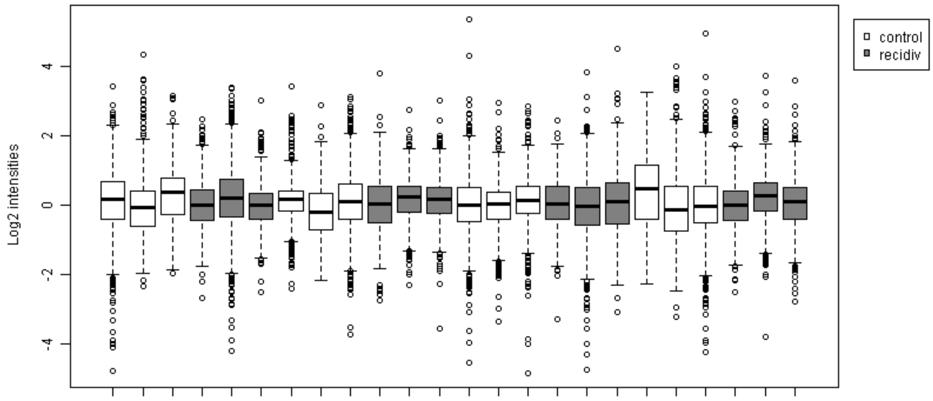

Figure 5. Example of a boxplot for protein expression data from two clinical sample groups: control and recidiv (relapse). The intensities of all proteins for one sample are plotted in each column. The horizontal line marks the median intensity for each sample, the boxes covers the mid 50% of data, spanning from the first quartile to the third quartile. The upper and lower vertical lines marks a 95% confidence interval for the median and the circles outside the lines are thus outliers.

1.4.2 Statistical analysis

In a large scale proteomics experiment we regularly start with thousands of variables, although we expect only a small fraction of those to be interesting for the biological or clinical question. The purpose of the statistical analysis is to extract the variables/proteins that are important for the clinical or biological question at task [113]. Most often the goal is to do a group comparison of samples from different conditions. Statistical methods for group comparisons can be divided into univariate methods, that test one variable at a time, and multivariate methods, that test all variables simultaneously.

1.4.2.1 Univariate methods

The most commonly used univariate method for group comparison is the student's t-test, which compares the distribution of a variable between two groups of samples. T-test exists for one or two groups. For more than two groups the ANOVA is an alternative. Both t-test and ANOVA are parametric which mean they rely on certain assumptions about the data – that it is normally distributed and has a homogenous variance (homoscedasticity). If those assumptions are not met, there exist non-parametric methods for both two group (Mann Whitney) and multiple group comparisons (Kruskal-Wallis).

1.4.2.2 Multiple testing problem

The significance level of a hypothesis test is often expressed in terms of p-values. The p-value indicates the probability that the relationship or difference found in the sample occurred by chance and is used to control the type I error (false positives) [112]. In proteomics studies, where many proteins are tested simultaneously, the probability of committing a type I error increases dramatically. The problem is that the standard hypothesis test is designed to control the type I error of each test at certain significance level. As the number of independent tests increase, the likelihood of observing data that satisfies the rejection criterion by chance alone increases (type I errors). To control the experiment-wise error rate, alternative measures of error are needed in those cases [112, 114]. Commonly used measures of error in multiple testing procedures are family wise error rate (FWER), the probability of at least one type I error, and false discovery rate (FDR), the expected proportion of type I errors among the declared significant results.

There are two main approaches for controlling the experiment wise error rate: Methods for controlling the FWER, like the Bonferroni method [115], and methods for controlling the FDR, like the Benjamini & Hochberg step down method [116]. Methods to control FWER are appropriate when you want to guard against any false positive. However, in many cases (particularly in omics discovery experiments) this is too conservative and a certain number of false positives can be tolerated. In these cases, the more relevant quantity to control is the false discovery rate (FDR). Furthermore, controlling FWER may lead to a very high rate of false negatives (type II error).

1.4.2.3 Multivariate methods

If the phenotype or biological process studied is thought to be effected by several variables/proteins in combination, a multivariate approach is often more appropriate than a univariate. The strength of multivariate methods is the possibility to define combinations of variables that maximizes the model predictive ability. Characteristics of proteomics experiments are thousands of variables (features) and small sample size. This is called the high-dimensional small-sample problem, which causes several statistical methods to fail or perform sub-optimal [113, 117]. Therefore, dimension reduction is often a necessary step in the analysis of proteomics data. Dimension reduction can be divided into feature selection and feature extraction (transformation). Feature selection methods reduce the number of features by excluding irrelevant or redundant features. Feature extraction methods identify a new set of features by transforming or combining the old features [117].

Commonly used multivariate methods for dimension reduction, variable selection and classification are Principal Component Analysis (PCA) and Partial Least Squares (PLS). PCA and PLS can handle high dimensionality of the data, as well as the presence of a large amount of biological noise. PCA is an unsupervised method [118], useful for getting an unbiased overview of the data as well as to detect trends and outliers. A PCA model is generated by introducing a new set of variables, which maximize the variance of a linear combination of the original predictor variables. The new variables, called principal components, represent directions in the data demonstrating the highest variation (Figure 6). This might of course be distinctly different from the directions best separating the classes. PLS regression is a supervised multivariate method for assessing the relationship between a descriptor matrix X and a response matrix Y [119, 120]. PLS takes the classes in the data into account and finds new variables by maximizing the covariance between the response variable and a linear combination of the predictor variables (see more details in section 2.2.8).

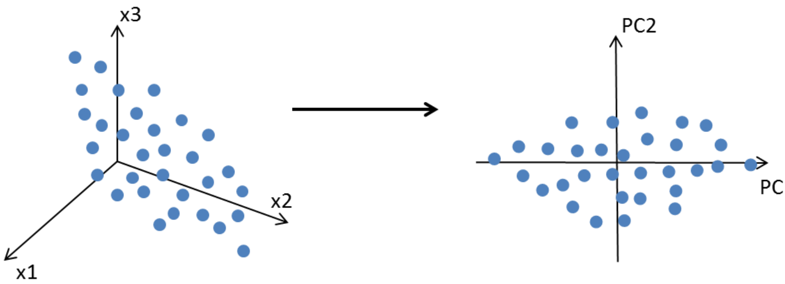

Figure 6. Schematic figure of principal component analysis for a simple case of three predictor variables (x1, x2, x3). The PCA model is generated by introducing a new set of variables, which maximize the variance of a linear combination of the original predictor variables. The new variables, called principal components (PC1 and PC2), represent directions in the data demonstrating the largest variation.

1.4.2.4 Model validation

Any statistical model needs to be validated, to assess the stability and generalizability of the model [113]. In an optimal scenario a completely new set of samples would be used to test the model performance on. In most cases, this is not possible due to few samples available. An alternative way to validate the model, which is commonly used, is cross-validation. In a *k*-fold cross-validation the data is randomly divided into *k* parts. The model is built and optimized on *k*-1 of the parts and tested on the excluded part. This is repeated for all the *k* parts and average model performance is calculated. During cross-validation it is important to remember to handle replicates together; otherwise one might risk receiving overly optimistic model performance.

1.5 BIOLOGICAL INTERPRETATION

The output of a statistical analysis of proteomics data is one or more lists of proteins that show an interesting change in level in the context of the experiment. This is not the end point of the analysis, but the starting point of a very complex process of deriving biological interpretation. The biological interpretation aims at placing the selected proteins into a context, to lift the analysis from individual molecules to the biological system level. During the biological interpretation process, the molecular expression data from the proteomics experiment is coupled with the vast information held in public knowledge databases [121, 122].

1.5.1 Systems Biology

To enable the leap from data analysis to biological interpretation, system based approaches integrating multiple data types are crucial. Systems biology is the study of systems of biological components, with the focus on complex interactions between the components [123-126]. Living systems are dynamic and complex and their behaviour are hard to predict from the properties of the individual parts.

Biological processes are often driven by modules of proteins working together rather than individual genes or proteins. Several comprehensive studies, mostly in cancer, have shown very few genes that have a robust and significant differential expression pattern across different sample cohorts [127, 128]. However, many of the sample cohorts showed similar differentially regulated pathways [129-132]. By moving the omics field from single molecules to affected pathways or network modules, we can generate models of the system which are more readily interpretable as well as more robust.

1.5.2 Networks and Pathways

Networks are built up by components (nodes) and interactions (edges) between them. The interaction can be almost any kind of association and can be directed or undirected. For example is a protein-protein interaction network built up by proteins (the nodes) and the physical interactions between them (the edges) [133]. Pathways are also networks, the difference lays in the level of annotation or understanding. Typically pathways are well-defined parts of the network that relates to a known physiological process or complete function, for example Glycolysis, Amino acid metabolism or Cell cycle. There are numerous databases available for networks and pathways. For a comprehensive listing of biological pathway and molecular interaction related resources see www.pathguide.org [134].

Biological networks have shown to be very different from random networks (randomly connected molecules); they apply to some basic organizing principles in their structure and evolution. For example, biological networks show a high degree of clustering and presence of a few highly connected nodes (hubs) that hold the network together [135]. By the use of networks and pathways we can integrate different types of molecular expression data and form modules of biologically related proteins. The network is the backbone, placing the molecules into a topological context. Molecular expression data on the other hand, has quantitative measurements of the molecules in a sample under different conditions. The integration of those different sources of information (i.e. expression data and networks) holds great potential to give new insights into disease biology [135-137].

1.5.2.1 Regulated subnetworks

Several methods have been developed to integrate expression data with interaction maps or pathway databases with the aim to identify subsets of the network (subnetworks) that associates with biological or clinical outcome. The subnetworks are sets of interacting proteins whose combined expression data can predict or classify samples. Numerous recent publications have shown that the predictive performance of expression data can be improved by the incorporation of interactome data [138-140]. Compared to traditional individual marker genes, the identified subnetwork markers had several advantages, as they more readily provide models of molecular processes and are more robust and predictive [138].

2 THE PRESENT STUDY

2.1 AIMS

The overall aim of this thesis work was to generate robust methods for selection of key proteins, networks and pathways relevant in relation to biological and clinical questions, using vast experimental proteomics data as starting point. This includes development and evaluation of methods for quantitative analysis of proteomics data, proceeding from setting adequate limits of quantification to statistical data analysis methods and system based approaches for integrating several types of data, towards the goal to generate biologically and clinically relevant information.

The specific aims of the papers I-IV were:

Paper I: To develop a multivariate meta-analysis workflow to couple 2DE data from colon and prostate human tumours, to identify common and unique protein patterns for the two tumour types.

Paper II: To develop a methodology for improved protein quantification in shotgun proteomics data and introduce a way to assess quantification errors for proteins in complex biological samples.

Paper III: To develop a tool for splice variant identification and visualization based on MS proteomics data, to provide the possibility to perform splice variant specific quantitative analysis.

Paper IV: To develop a multivariate network based analysis workflow for proteomics data to identify subnetworks with different activity between groups of samples, to enable detection of differences on a biological system level and to further enhance the interpretation of results from cancer studies.

2.2 MATERIALS AND METHODS

This section describes some selected key methods and aspects applied in **papers I-IV**. The materials and methods are described in detail in each paper.

2.2.1 Samples and Study design

Proteomics data from both cancer cell lines and tumour material was used in the present study. **Paper I** includes 2DE data on samples from human prostate and colon tumours. The approach in **paper II** was first evaluated on a standard dataset of A549 cell lysate mixed in the proportions 2:2:1:1:2:2:1:1 (see experimental setup in Figure 7). To demonstrate the usability, the methodology was also applied to another cancer cell line experiment as well as in a clinical dataset of lung cancer tissue samples. To exemplify the capability of the software developed in **paper III**, the method was applied on an experimental dataset of A431 cell line treated with Gefitinib. The analysis developed in **paper IV** was tested on different complex biological datasets, both cell line samples and clinical samples.

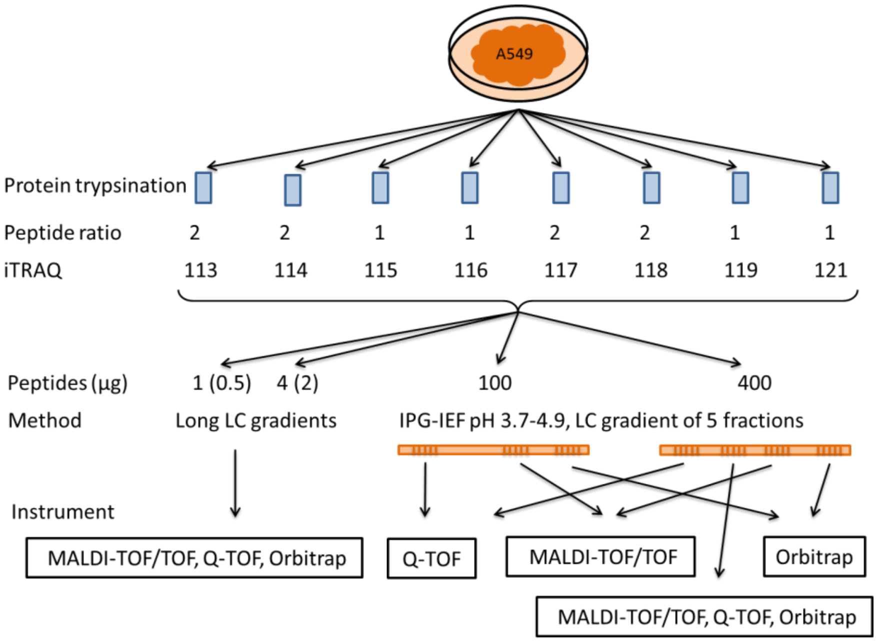

Figure 7. Experimental setup for standard dataset in paper II. Tryptic peptides from A549 cells were labelled with iTRAQ in a 2:2:1:1:2:2:1:1 ratio. Peptides were analysed by LC-MS/MS alone or pre-fractionated before LC-MS/MS using narrow range immobilized pH gradient isoelectric focusing (IPG-IEF). A mix of all peptides or extracted peptide fractions from the IPG-IEF were analysed on three different LC-MS platforms.

2.2.2 Two-dimensional gel electrophoresis

In **paper I**, data from two-dimensional gel electrophoresis (2DE) on prostate and colon tumours were subject to a multivariate meta-analysis. The proteins were separated in the first dimension of isoelectric focusing using immobilized pH-gradient (IPG) strips with a pH 4-7 linear gradient. The second dimension was performed using 10-13% linear gradient SDS/PAGE gels. The gels were then stained and scanned and the images analysed by the PDQuest software [141]. The two sample sets, prostate and colon, were first analysed separately in the software. The masters (image containing the most spots) from the separate match sets were then matched to each other and thereby linked all the gels in the two data sets together.

2.2.3 Isoelectric focusing

In **papers II**, **III** and **IV** narrow range immobilized pH gradient isoelectric focusing (IPG-IEF) was used on peptide level to reduce sample complexity [142, 143]. On an IPG-IEF strip, the peptides are separated according to their isoelectric point. The complexity of the peptide mixture is thereby reduced by selectively analysing the sub-fraction of peptides within a certain p*I* range. Different pH ranges can be used dependent on which fraction of the peptidome one would like to focus on. The acidic pH range (3.7-4.9) in these studies is chosen so that the complexity is reduced without any significant loss of proteome coverage [143]. The p*I* of identified peptides can further be used to validate the peptide sequence and to restrict the search database [144, 145].

2.2.4 iTRAQ and TMT labelling

Papers II, **III** and **IV** uses 8-plex iTRAQ based quantification of peptides. iTRAQ (isobaric tags for relative and absolute quantification) allow for quantification of up to eight samples within a single run. Using iTRAQ, fragmented reporter ions from the tag are used for quantification in MS/MS mode. The intact iTRAQ labels have the same mass and same MS properties. The individual tags are distinguished by their fragmentation patterns in MS/MS, giving rise to reporter ions of different masses that can be quantified in the MS/MS spectra. iTRAQ is primarily used for relative quantification, the ratio between the reporter ions within one spectrum is used for relative quantification of each peptide within one run. If more than eight samples are analysed, comparison between runs is necessary, for this one commonly uses an internal standard shared between the runs. **Paper IV** also includes one dataset with quantification by tandem mass tags (TMT). TMT is similar to iTRAQ, but the reporter ions have slightly larger mass and exist in six tags.

2.2.5 Mass spectrometry instruments

MS instruments consist of three major modules: an ion source, a mass analyser and a detector. Each type of MS instrument uses a different setup of those three modules. In **papers II**, **III** and **IV** an LTQ-Orbitrap [146, 147] was used for LC-MS/MS analysis. The LTQ-Orbitrap was coupled to a nano-ESI source that ionizes the peptides eluting from the LC column. LTQ-Orbitrap Velos [148] is a hybrid instrument with two different kind of mass analysers: a LTQ (Linear Quadrupole Ion trap) which separates ions based on stability of their paths in oscillating electric fields and an Orbitrap that separates ions based on m/z resonance frequency. This LTQ-Orbitrap thus combines the sensitivity and speed of the LTQ with the high mass accuracy and high resolution of the Orbitrap [146, 149].

In the LTQ-Orbitrap Velos, the peptide ion mass spectrum (MS1) is acquired with the Orbitrap. The fragmentation of the peptide ions can be done either in the ion trap, using collision induced dissociation (CID), or in the higher-energy collisional dissociation (HCD) chamber. The fragment ion mass spectrum (MS2) can then be acquired either in the ion trap or in the Orbitrap. The low energy CID fragmentation results in an escape of many small ions (low mass range) leading to low quality spectra in the low mass region, where the iTRAQ reporter ions end up. Using the higher energy collision (HCD), the small ions stay and the quality of the spectra in the low mass region is good, thus the reporter ions can be used for quantification [150]. Optimal settings for the LTQ-Orbitrap Velos, regarding collision energy for HCD and fragmentation time, are investigated in **paper II**.

In **paper II**, the performance of several different MS platforms were compared and data was also generated on a MALDI-TOF/TOF [151] and Q-TOF [152] system. Matrix assisted laser desorption ionization (MALDI) is a soft ionization method used in mass spectrometry. In MALDI the sample co-crystallizes with a matrix and is pulsed with a laser, which ionizes and vaporizes the analytes. The MALDI ion source is most often coupled to a time-of-flight (TOF) analyser, which uses flight time to separate ions. TOF/TOF is a tandem mass spectrometry method where two time-of-flight mass spectrometers are used consecutively to generate MS2 spectra. The Q-TOF is another hybrid instrument with a Quadrupole coupled to a time-of-flight analyser. Each instrument has its own advantages as well as disadvantages, and is suitable for different types of studies [35, 64].

2.2.6 Peptide and Protein identification

2.2.6.1 Database search

The output from tandem mass spectrometry analysis is precursor (peptide) ion spectra (MS1) and fragment ion spectra (MS2). The fragment ion spectra need to be assigned to peptide sequences to be able to infer which peptides, and thereby proteins, were present in the sample. The method for peptide identification used in **papers II**, **III** and **IV** was database search [153]. The database consists of all protein sequences downloaded from for example Ensembl (www.ensemble.org) [154]. The protein sequence is then theoretically digested by trypsin to generate peptide sequences. The database is then searched, using a search engine, for the peptide whose predicted spectrum best matches the observed spectrum. To limit the possible matches in a database search, the search is restricted by mass tolerance, proteolytic enzyme constraints, post-translational modifications and the m/z of the precursor ion (peptide).

The output from a database search is a collection of peptide-spectrum matches (PSM) with an associated score. The score reflects the similarity between measured and predicted spectra. A number of different search algorithms and scoring schemes have been described in the literature [94, 95, 97], commonly used publically available tools are Mascot (used in **paper II**) and Sequest (used in **paper III** and **IV)**.

The peptide score depends on dataset, search algorithm and search parameters, which makes it very difficult to compare scores between search algorithms and datasets. Methods have been developed to provide statistical measures of confidence and estimates of error rate, which are independent of the scoring scheme used [94]. The statistical approaches can be grouped into two categories: target-decoy approaches [155] and empirical Bayes approaches [156, 157]. The target-decoy approach was used in **papers II**, **III** and **IV**. This approach is based on creating a decoy database, which is a reversed or shuffled version of the target database, and then search the two databases with the same settings. Assuming that there is no overlap between the target and decoy databases and assuming that incorrect assignments from target and decoy sequences are equally likely, we do not expect to get any real matches from a decoy database. The number of matches found in the decoy database is thus a good estimate of the number of false positives present in the matches from the target database. The target-decoy approach gives robust and effective estimates of the number of incorrect identifications (FDR) for an entire dataset, but it does not remove incorrect identifications. With the use of the target-decoy approach one can select the score threshold needed to reach a certain FDR level (1% used in **papers II-IV**).

Most often, the peptides identified by the database search are grouped into proteins, and one would therefore like to control the FDR at the protein level. But since errors determined at the PSM level, by target-decoy approach for example, propagate to protein identification level in a non-trivial manner, this is not a straight-forward task. One method for computing FDR at protein level is MAYU [158], which was used in **paper II**.

Using a database search, only around 25-30% of the generated spectra can be explained successfully. The unexplained spectra can have several reasons, like for example poor quality spectra. One other big limitation is incomplete databases not containing all the protein variants present in the sample. This is specifically true for protein isoforms, and post translational modified proteins, which often are poorly covered in the traditional databases and search engines.

2.2.6.2 Inferring protein

Protein identifications are defined as assemblies of PSMs. The protein inference, which groups peptides into proteins, faces many challenges: many peptides group into relatively small number of proteins, incorrect spectral identifications match randomly to the large protein database and shared peptides make it difficult to separate out protein isoforms. Alternative splicing is a widespread event in the human genome, as much as around 90% of human genes undergo alternative splicing. Splice variants share peptide sequences to a large extent and is therefore difficult to separate out by database search.

One method to try to increase the number of splice variants detected by mass spectrometry has been to include the sequence of known and predicted protein variants in the search database [159]. However, this method expands the searching space significantly, effecting searching time and risk of making false peptide discoveries. In **paper III,** an alternative method for identifying and quantifying splice variants in mass spectrometry based proteomics data is developed. The developed tool mines data from the alternative splicing database EVDB (Evidence Viewer Database) [160] and maps MS identified peptides to known splice variants.

2.2.7 Peptide and Protein quantification

In the current study (**papers II-IV**), the iTRAQ reporter ions are used for relative quantification of the peptides in each of the eight samples. The peptide ratios are calculated by dividing each iTRAQ channel by the mean of the first two iTRAQ channels (113, 114). In **papers II-IV,** the peptide ratios are normalized to the same sample median on peptide level to make iTRAQ channels comparable, assuming that the peptide distribution is equal between samples. This assumption is also based on the fact that the protein amount loaded is equal for all samples. The protein ratios are also log2 transformed to bring low signals and high signals more together and to make up and down regulations equally important.

The peptide ratios are then aggregated to yield protein ratios. The quantitative measurements on the peptide level have to be aggregated to protein quantification in a way that returns the best (most accurate and precise) protein quantification measure. Most methods for summarizing peptide data into protein data rely on a simple mean or median over the peptide ratios. By this method, low intensity signals or noisy data as well as wrongly assigned peptides may easily distort the computed protein ratios. A recent paper introduced a novel statistical estimator for protein ratios, generating improved protein quantification as well as a built-in quality control metric [161]. In **paper II**, some methods for summarizing peptide data into protein ratios are compared. The presence of several protein isoforms in the sample can potentially also cause incorrect protein ratios. If several unresolved protein isoforms are present, the protein ratio is a mixture of different protein species. Recently, a tool for Protein Quantification by Peptide Quality control (PQPQ), was developed [162]. PQPQ looks at the correlation pattern for peptides over iTRAQ channels to detect peptide clusters and outlying peptides and includes only correlated peptides in the calculation of the protein quantities. **Paper III** investigates the effect of unresolved protein isoforms by comparing the quantification based on gene centric, protein centric and splice variant centric analysis.

In iTRAQ, systematic biases can arise because of differences in iTRAQ labelling efficiency and protein digestion. Recent studies have reported that iTRAQ data has issues with both accuracy and precision [163, 164]. Fold changes were underestimated and biased towards null. The precision was affected by variance heterogeneity, with higher variance for low intensity signals. This is a problem since low signals dominate the data sets and many proteins have only few peptide readings. Improved quantitation methods have been suggested, addressing the variance heterogeneity by excluding low intensity peptides [165], weighting peptide data by uncertainty [165-168] or stabilizing the variance [163]. In **paper II**, the errors in iTRAQ quantification is investigated and an improved method for protein quantification is suggested.

2.2.8 Statistical analysis

The statistical method is used to prioritize or select the proteins believed to be important for the clinical or biological question. From 5000-10 000 proteins identified in a high quality shotgun proteomics experiment, the statistical analysis narrows down to a few (10-100) proteins of interest. In **papers I** and **IV**, a multivariate PLS model is used as the basis for selection and evaluation of proteins. PLS regression is a multivariate method for assessing a relationship between a descriptor matrix X and a response matrix Y. In the context of the proteomics data, the protein expression data is the descriptor matrix and the sample class is the response matrix.

PLS models are generated by finding latent variables (PLS components) in the data that maximize the covariance between the response variable (Y) and a linear combination of the predictor variables (X). PLS Discriminant Analysis (PLS-DA) is a classical PLS regression but where the response variable is categorical, indicating the classes of the samples. PLS-DA has often been used for classification and discrimination problems [169, 170]. An extension to the supervised PLS regression method is Orthogonal projection to latent structures (OPLS) [171]. OPLS uses information in the Y matrix to separate the X matrix into correlated (predictive) and uncorrelated (non-predictive) orthogonal information. Those changes often lead to an improved interpretability, while the predictivity is the same as for the PLS model.

PLS models can be used for regression, classification, prediction and variable selection. The strength of PLS lies in the interpretation of the model and the variables importance for the model. The usage of PLS in the current study has been mainly to select proteins of interest. For this the Variable Importance on Projection (VIP) was used, which is a summary of the importance of a variable on the model [120].

2.2.8.1 Statistical model validation

The PLS models are validated by cross-validation in **papers I** and **IV**. In **paper I**, a double cross-validation scheme is used [172]. The inner loop consisted of a bootstrap cross-validation [173] for the optimization of PLS model parameters and variable selection. The outer loop was a 5-fold cross-validation used to evaluate the performance of the optimal model. The variable selection was based on mean VIP score as well as stability over cross-validation rounds. In **paper IV**, the subnetwork PLS model is evaluated based on a leave-one-out (LOO) cross-validation. In general, LOO should be used with care since leaving out only one sample might lead to over-optimistic results caused by other similar samples in the training set. The choice of LOO for the subnetwork model was based on that one of the proteomics datasets consisted of very few samples (3+3).

The model performance can be assessed by several different measures. R2 and Q2 are commonly used for multivariate PLS and OPLS models. R2 is a measure of how well the model describes the data, thus it is based on the training dataset. R2X is the fraction of X variance explained by the model, while R2Y is the fraction of Y variance explained by the model. Q2 is based on the test set during cross-validation and is a measure of how well the model predicts “new” data. The subnetwork PLS model in **paper IV** was evaluated based on Q2. Other frequently used measures of model performance are sensitivity and specificity, which are also calculated from the prediction of a test set. Sensitivity, or true positive rate, is the probability of a positive test among positive samples. Specificity, or true negative rate, is the probability of a negative test among negative samples. The success measure used to evaluate the PLS model in **paper I** was the geometric mean of sensitivity and specificity.

2.2.8.2 Subnetwork methods

In **paper IV**, protein expression values from MS analysis were mapped to protein interaction data. A network based PLS model was then developed to identify subnetworks differentially regulated between phenotypes. Several different approaches to identify differently regulated subnetworks, based on expression data, have been published over the last years. Most methods have two components in common: a scoring method to measure the discriminative strength of the subnetwork and a search algorithm to find the highest scoring subnetworks. The scoring methods used have been basic scoring schemes such as absolute difference [140], p-values [174] and mutual information [138], as well as more complicated scoring strategies based on principal components [175], decision trees [176] and support vector machines [177].

The scoring approaches can roughly be divided into univariate and multivariate [139]. A univariate scoring approach assesses the regulation of each node individually and then searches for subnetworks with enrichment in regulated nodes. A multivariate scoring approach on the other hand, assesses regulation for all nodes in the subnetwork together. The scoring in **paper IV** was based on a multivariate PLS model, evaluated by Q2 based on a LOO cross-validation on the samples. The possible variables in the PLS model were thus restricted by the links in the network. A greedy search algorithm [138] was used to search for the optimal scoring subnetwork initiated from each starting node.

The generated subnetworks each have a score based on the Q2 of the PLS model. The significance level of the scores has to be assessed by randomization. Random networks were created by node permutation [178] in **paper IV** to define a score threshold for significant subnetworks.

2.2.8.3 Network data

To increase the protein network size in **paper IV**, a meta database (STRING) of protein associations was used. STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is a resource for retrieving all information on functional links between proteins [179, 180]. The associations in STRING are derived from high-throughput experimental data, from text-mining of databases and literature, and from predictions based on genomic context analysis. STRING combines and scores interaction data from the various sources for a large number of organisms, and also transfers information between the organisms via orthologous protein pairs. The confidence score for each link, reflects how likely a given association is. The database currently covers over 5 million proteins from 1133 organisms (Version 9.1). For **paper IV**, the network was restricted to only Human interactions with a confidence score higher than 900, considered to be highly confident (Figure 8).

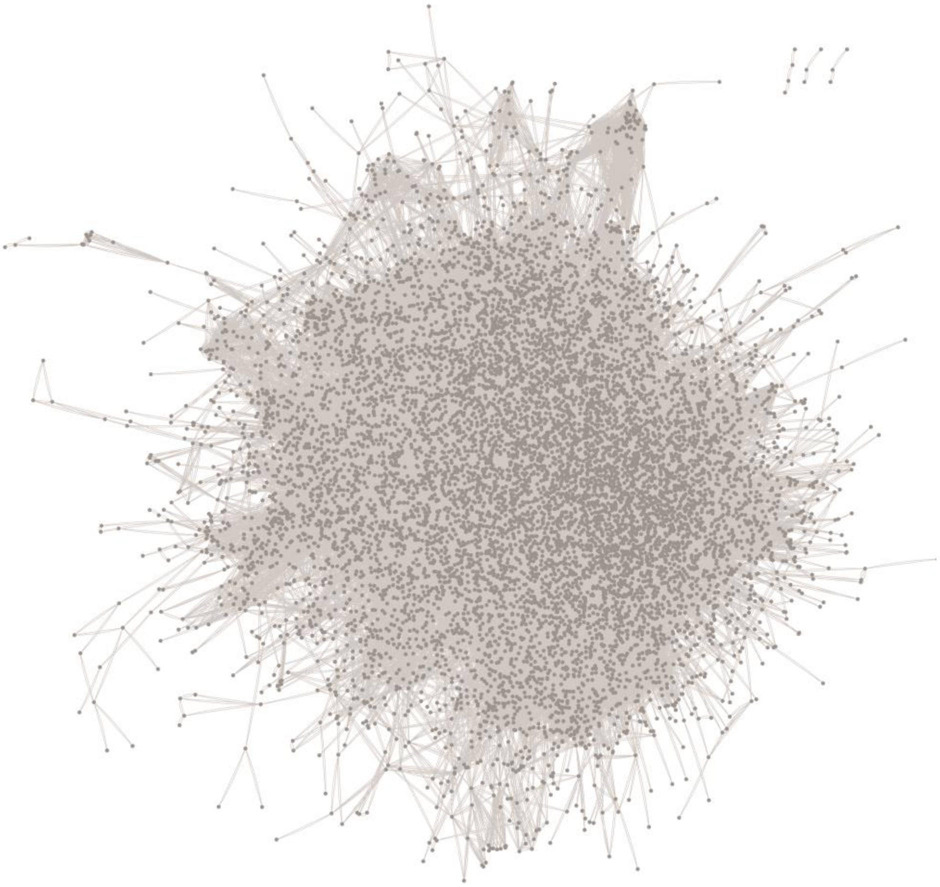

Figure 8. Network of protein interactions from STRING. The nodes are proteins and the edges between them are functional couplings. The network is restricted to only Human interactions with a confidence score higher than 900. The network consists of 113306 interactions and 9542 proteins.

2.3 RESULTS AND DISCUSSION

2.3.1 Paper I

In this work, we demonstrated a multivariate meta-analysis of 2DE proteomics data from human prostate and colon tumours, with the aim to identify common and unique protein patterns. The bioinformatics workflow developed included merging of the two datasets followed by dealing with pre-processing of data and handling of missing values and the development of a multivariate PLS model for prediction and variable selection. The missing values posed a big challenge in analysing the 2DE data from two very different tumour types. Many missing values existed in the merged data set, which affected the modelling result. With the purpose of finding proteins with common expression patterns over the two tumour types, the analysis was restricted to those proteins that were expressed in both data sets.

This study utilized PLS-DA to build predictive models and to select variables important for separating between the classes normal and tumour, independent of tissue origin. The PLS model development and variable selection was rigorously evaluated using a double cross-validation scheme (Figure 9). The mean success rate over bootstrap rounds in the outer loop was plotted for varying number of variables and number of PLS components and used to find an optimal PLS model.

The optimal number of variables and PLS components in the PLS model is a trade-off. The number of selected variables should be small enough to enable further validation of the proteins using more targeted methods for measuring the expression levels in a larger cohort of clinical samples. At the same time, the number of variables has to be large enough to achieve a good predictive PLS model. Regarding the PLS components, too few components might not be enough to explain the data while too many might lead to an over-fitted model, describing the noise in the data. For the current study, three PLS components and 50 variables were selected as optimal PLS parameter settings.

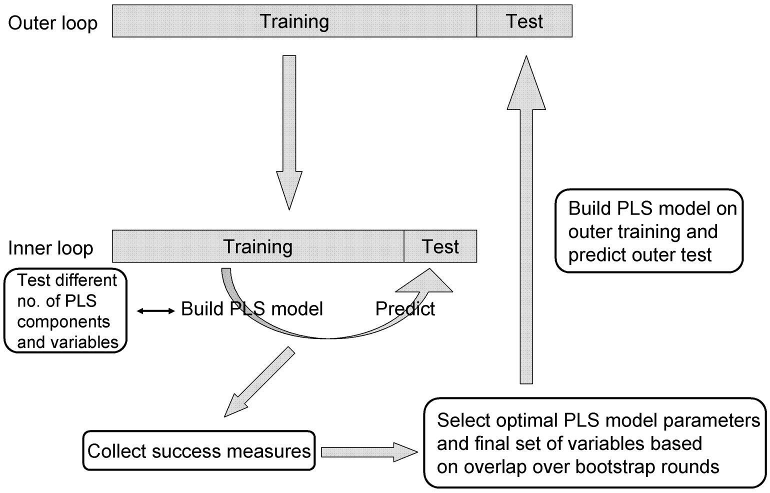
Outer loop
Training
Test
Build PLS model on outer training and predict outer test
Inner loop
Training
Test
Test different no. of PLS components and variables
Build PLS model
Predict
Collect success measures
Select optimal PLS model parameters and final set of variables based on overlap over bootstrap rounds

Figure 9. Double cross-validation scheme. In the inner loop, PLS model parameters and variables are estimated based on a bootstrap cross-validation. Based on performance of the PLS models and stability of variables over bootstrap rounds, the optimal parameters and final set of variables are selected. Model performance of the optimized parameters and selected variables are then evaluated on the held-out test set in the outer loop. The outer loop is repeated within a 5-fold cross-validation procedure.

The final selection of variables was based on stability over the bootstrap validation rounds in the inner loop. The reasoning is that the stable variables are thought to represent variables generally good for predicting the classes and not specific for certain subsets of the data. Despite such different tissues in the data, there were around 40 variables (from the lists of 50 variables) that were selected in at least 50% of the bootstrap rounds. The stable variables were together with the optimized PLS model applied to predict the held-out test sets in the outer loop. The average prediction success over five cross-validation rounds was 0.93 (±0.06), for the PLS model discriminating between normal and tumour samples, independent on tissue type.

The combined prostate-colon model was compared to individual prostate and colon PLS models (including only variables present in both datasets). The resulting lists of stable variables for the three models were compared in a Venn diagram (Figure 10). The figure reveals that most variables are unique to the models and few overlaps are identified, only three variables overlap between all three models. As many as 46 of the variables from the meta-analysis of prostate-colon did not show up in the individual models, while 25 and 27 variables were unique to the prostate and colon models respectively. The 46 variables unique to the meta-model represent proteins whose expression levels discriminate between normal and tumour samples independent of tissue type in this study, i.e. a common protein profile for malign tumour types. The variables unique to the individual models on the other hand represent proteins that are specific for the certain tumour types prostate and colon. This result shows the potential of a meta-analysis to identify proteins not found when analysing the data sets in separate. The current study only included two tumour types, and can mainly function as a proof of concept, but the potential of including more tumour types is apparent.

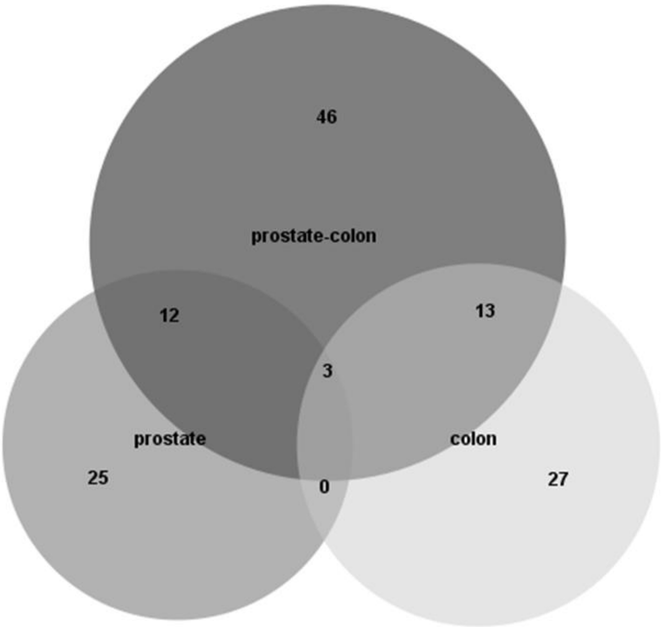

Figure 10. Overlap of variables selected in different models. Venn diagram showing overlap between stable variables selected using the prostate-colon meta-model and individual prostate and colon models.

2.3.2 Paper II

In this study, we developed a methodology for improved protein quantification in shotgun proteomics data and introduced a way to assess quantification quality. Peptide and protein identification and quantification was compared between different MS platforms, as well as between different loaded peptide amounts and different sample separation methods. See experimental setup for the standard dataset in Figure 7.

The quality of the peptide quantification was evaluated by scaled root mean square error ($RMSE_s$). The $RMSE_s$ includes both bias and variance and measures the average magnitude of the error per peptide over all eight iTRAQ channels. The $RMSE_s$ values were plotted against intensity, revealing that the error in quantitation is intensity dependent and decreases as the peptide intensity increase. To be able to study only the variance in the peptide quantifications, the peptide intensities were normalized to equal sample median and the relative standard deviation (RSD) calculated. RSD and $RMSE_s$ shows the same trend with decreasing RSD when intensity increases. The RSD was overall smaller than $RMSE_s$ showing that there is a bias in the un-normalized data. Further investigations exhibited a small bias (around 5%) towards one. In our settings, the variance thus seems to be the largest contributor to the error.

$RMSE_s$ was calculated to compare instruments, loaded peptide amount and separation method. The peptide quantities from the Orbitrap and MALDI have rather similar $RMSE_s$ values, while QTOF peptide quantities have much higher $RMSE_s$ values. The number of identified peptides also varied largely with the MS instrument, the Orbitrap generated more than five times as many identifications as the MALDI and QTOF. The results on protein level mainly confirm the results from the comparison on the peptide level; the Orbitrap performs best followed by MALDI and then QTOF. Orbitrap identifies approximately four times more proteins than the other instruments do. In summary, increasing the amount of loaded peptides as well as pre-fractionating the sample by IPG-IEF results in the best performance for the Orbitrap, both when it comes to error levels at the peptide and protein level as well as number of identifications.

It is crucial that the quantitative information on the peptide level is correct when summarizing to protein level quantity. In this study we therefore evaluated two alternative methods to improve protein quantities: either by removing low intensity peptides prior to summarizing to protein quantity or by using all peptides but weight them according to their uncertainty (determined by their absolute intensity, high weight corresponding to high intensity) when summarizing to protein quantity. The weighted mean and filter methods were compared to using all peptides for the calculation of a regular mean as well as to the weighted mean method in the Mascot software. The measured protein ratios were compared to the expected ratios and the relative error was calculated for all protein quantification approaches (Figure 11A). It can be seen in the figure that more proteins are calculated with a lower relative error when using the weighted mean as compared to the other methods.

In Figure 11B, the relative error of protein quantity is related to protein weight (calculated as the mean of the peptide weights derived from the intensity) for proteins with different number of peptides. Seen in the figure, the relative error of the protein quantity is very much dependent on the number of peptides used for quantification of the protein. For proteins with few peptides, the intensity of the peptides (visualized by protein weight) influence the relative error strongly, while for proteins with large number of peptides the intensity of the peptides has smaller impact on error. Even at low protein weight the relative error is rather small for proteins with multiple peptides for quantification. Hence, peptides with low intensity can be important for creating a robust protein quantity.

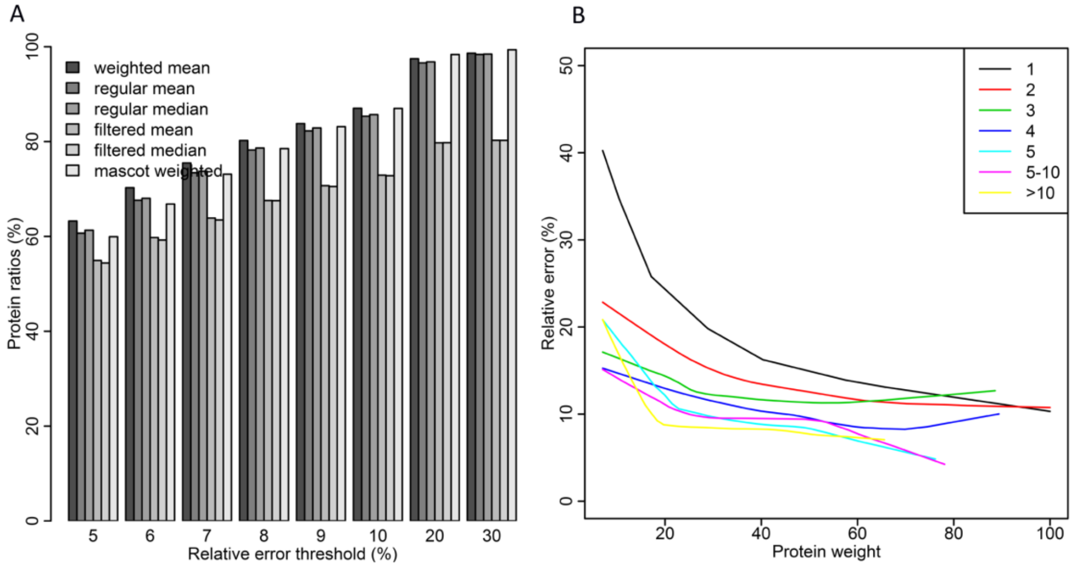

Figure 11. A: Comparison of methods to calculate protein quantities based on peptides. The bars represent percentage of protein ratios passing different relative error thresholds, for weighted protein mean, regular protein mean/median, filtered protein mean/median and Mascot weighted protein mean. B: Impact of the number of peptides per protein on quantification error. The relative error of weighted protein quantity is plotted against protein weight for proteins with different number of peptides. Lines represent smoothed 95% upper limit of relative error. The protein weight is calculated as the mean of peptide weights.

In the current study, the weight is calculated based on an internal training set (technical duplicate) for each run. An internal training set for the weights is to prefer, according to our results, since different experimental settings will affect the data quality differently. As an outcome of these results, we suggest including one technical duplicate in each iTRAQ run so weights can be calculated specifically for every new data set, and then be applied to the remaining biological iTRAQ samples. We further suggest that a plot (11B) and corresponding table with weights and errors are created for each dataset based on the duplicate in the experiment. This can then be used to set a threshold on protein weights to ascertain reliable protein ratios, which will be especially important for proteins with one or a few peptides for quantification.

The method of calculating weights based on an internal training set was applied to independent datasets of cell line samples and lung cancer tissue samples. The weighted mean performs slightly better than the regular mean, confirming the results from the original standard dataset. The improvement is rather modest, around 5% for the lung cancer tissue samples. Still, we believe this is an important improvement, as it corresponds to around 90 more proteins in the clinical dataset with accurate quantification (<5% relative error), which can be essential for discovering biomarkers.

2.3.3 Paper III

In this study we developed SpliceVista, a tool for splice variant identification and visualization based on MS proteomics data. By mining data from an alternative splicing database (EVDB) and mapping MS identified peptides to known splice variants, SpliceVista can identify splice variant specific peptides and perform splice variant specific quantitative analysis.

There are four main parts of SpliceVista: *Data pre-process*, in which all PSMs are assigned a gene symbol from its protein ID and grouped into peptides. *Download*, where SpliceVista retrieves known splice variants in the EVDB database and translated sequences of these splice variants from GenBank. *Mapping*, in which all identified peptides are grouped by gene based on the downloaded data. In particular, for each gene all the identified peptides are mapped to the gene's known splice variants. Genomic and transcriptional position of each peptide is reported in the output file. *Visualization*, where the data from previous steps is used for visualizing the exon structures of each splice variant of the protein and the transcriptional position of identified peptides. In addition, if PQPQ [162] is used, the peptide clusters based on quantitative information are visualized allowing connection between splice specific peptides and detected quantitative peptide clusters. See example in Figure 12.

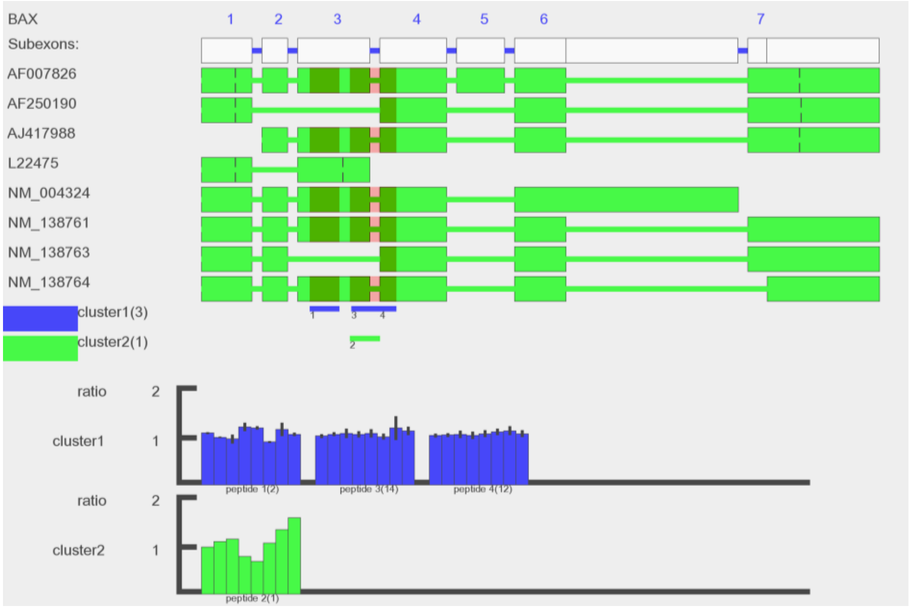

Figure 12. Example output figure of SpliceVista. The top panel displays the exon structure of the gene. The mid panel displays the transcriptional positions of identified peptides. If PQPQ is applied, each peptide is assigned to a cluster in which all peptides show correlated quantitative pattern. The different clusters are coloured and the peptides are coloured accordingly and plotted in line with the cluster it belongs to. In the bottom panel, the quantitative patterns of the different clusters are drawn in the same order as in the mid panel. The bars represent the mean intensity ratio of all peptide spectra matches (PSMs) for each unique peptide, the standard deviation is indicated by vertical lines (error bars).

To evaluate the potential and limitations of shotgun MS based proteomics for splice variant specific analysis, we performed in silico trypsin digestion of the human proteome (Ensembl 63). 18% of the tryptic peptides uniquely map to a splice variant and 22% of the splice variants have unique tryptic peptides. Given that a splice variant is present in the sample, the identification by shotgun proteomics is dependent on mainly two factors. First it depends on whether or not the splice variant has unique sequences to make it possible to identify. According to the theoretical calculation, up to 22% of human splice variants can thus be identified in theory by peptide centric MS by assigning splice variant specific peptides (SVSP). Secondly, it depends on the protein sequence coverage in the MS experiment. The higher the protein sequence coverage is, the higher the chance of identifying a splice variant by its unique peptides.

To test the applicability of the method on proteomics data generated by shotgun MS, we used SpliceVista to analyse human cancer cell line that had been analysed both as whole cell lysate, as well as through sub-cellular fractionation. 607 splice variants and 1680 SVSPs were identified in the whole cell fraction. After combining splice variants identified in the three subcellular fractions, the number of unique splice variants identified was 939 and the number of SVSPs was 2983. By subcellular fractionation, the number of splice variants and SVSP identifications were increased by 55% and 78% respectively. Theoretically, the chance of identifying one splice variant specific peptide is higher if there are more peptides per protein identified. As expected, the data demonstrated that using subcellular fractionation, we can increase splice variant specific peptide and splice variant identifications due to increased protein coverage.

We performed three different quantitative analyses on the genes with splice variants identified in the cell line dataset: gene centric, protein centric and splice variant specific analysis (Figure 13). In the gene centric analysis, the relative expression level of a gene is calculated by the mean ratio of all PSMs identified for this gene. In protein centric analysis, the conventional way, the relative expression level of a protein is calculated by the mean ratio of all PSMs for this protein. In splice variant specific analysis, PSMs specific (uniquely mapped) to one splice variant are grouped and the relative expression level of the splice variant is calculated as the mean ratio of those PSMs only.

The genes in Figure 13 exemplify cases where there is a large difference between gene centric, protein centric and splice variant specific analysis. Since more than 90% of genes can undergo alternative splicing, there is a potential risk of averaging out the differences of differentially regulated splice variants when doing protein centric analysis if the protein contains peptides shared among protein isoforms. With SpliceVista, we are able to quantify splice variants specifically and compare to gene centric and protein centric analysis.

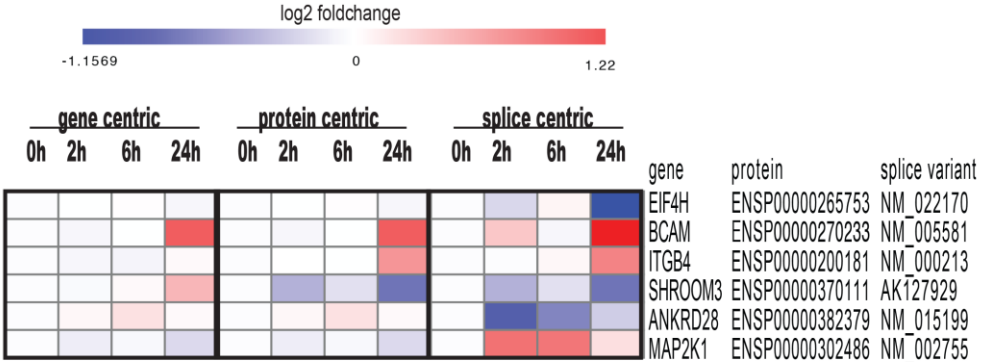

Figure 13. Heat map showing comparison of fold change between gene centric, protein centric and splice variant specific analysis at different time points.

2.3.4 Paper IV

In this work, we developed a network based analysis workflow for proteomics data to identify subnetworks with different activity between groups of samples. The idea was to shift focus from individual proteins showing differential expression to protein subnetworks with altered activity.

The outline of the subnetwork method is shown in Figure 14. The network data was extracted from the STRING database. Each of the proteins mapped to the network data was used as a starting node for the search algorithm, searching for the optimal scoring subnetwork. The search stops if any of the termination criteria are met; addition of neighbours does not improve the score over a defined threshold, 5%, or the size of the subnetwork is larger than 20 (to keep the search local), or if the protein node lacks additional neighbours. The resulting subnetwork scores were shown to be dependent on network size and the score distribution is thus not homogeneous.

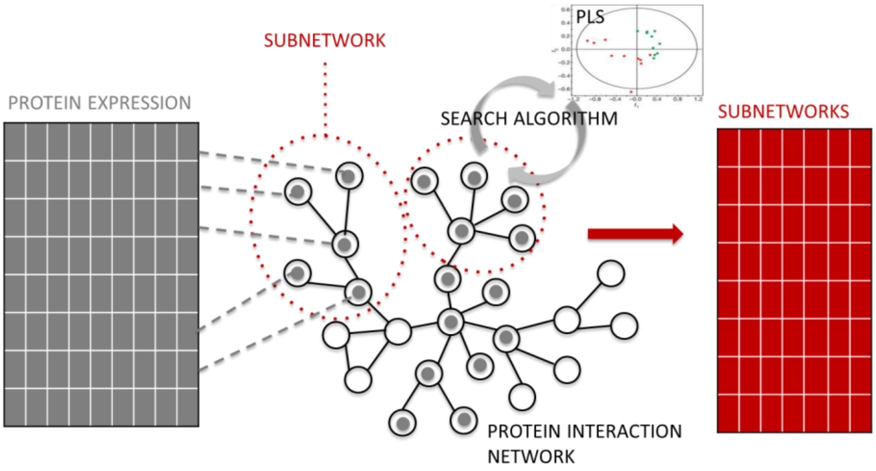

Figure 14. Schematic outline of the subnetwork analysis. The protein expression data is mapped onto the protein interaction network from STRING. A greedy search algorithm is searching through the mapped network for the highest scoring subnetwork. The score is based on predictive success (Q2) of a multivariate PLS model, evaluated by leave-one-out cross-validation on the samples. The optimal subnetworks and corresponding measurements of subnetwork score and size are saved in a new data matrix.

For evaluating the significance of optimal subnetworks, the results were compared to randomized input data. The network randomization was done by permuting the nodes 500 times, searching for optimal subnetworks and scoring them by the PLS model. Score threshold for any given significance level could then be calculated based on the fraction of random subnetworks exceeding certain score threshold. As was seen also for the real network, the score is dependent on subnetwork size, with a bias towards higher score for larger subnetworks. Since the score is clearly dependent on subnetwork size, the score threshold (corresponding to 5% FDR) was calculated for each subnetwork size separately. The resulting significant subnetworks were merged into one if they overlapped with at least two proteins. The largest significant subnetwork for the clinical data is seen in Figure 15.

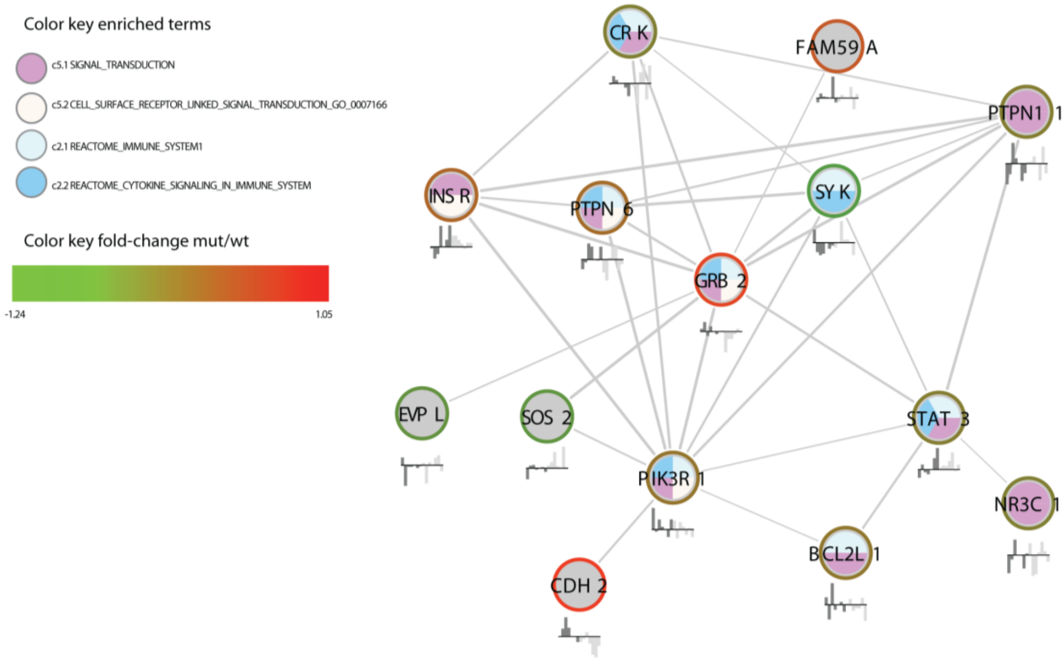
Color key enriched terms
c5.1 SIGNAL_TRANSDUCTION
c5.2 CELL_SURFACE_RECEPTOR_LINKED_SIGNAL_TRANSDUCTION_GO_0007166
c2.1 REACTOME_IMMUNE_SYSTEM1
c2.2 REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM
Color key fold-change mut/wt
-1.24
1.05
CR K
FAM59 A
PTPN1 1
INS R
PTPN 6
SY K
GRB 2
EVP L
SOS 2
STAT 3
PIK3R 1
NR3C 1
BCL2L 1
CDH 2

Figure 15. The figure depicts the largest significant subnetwork for the clinical dataset. The border of the nodes are coloured by fold-change of mutated versus wild-type, red is up-regulated in mutated and green is down-regulated in mutated. The individual log2 expression values are shown as a bargraph under each node, the mutated samples as dark grey bars and the wild-type samples as light grey bars. The four most enriched terms for the subnetwork are illustrated as a pie chart in each node, grey means that none of the terms are annotated to this node. The figure was generated in Cytoscape, an open source software platform for network visualisation [181].

The significant subnetworks were used as a basis for enrichment analysis, using gene sets from the MSigDB (Molecular Signatures Database) [182]. MSigDB is a collection of annotated gene sets, derived from a number of different databases as well as from computational approaches. Hypergeometric testing adjusted for multiple tests was performed to identify enriched terms. Only terms with at least three hits in the subnetwork and enriched by an adjusted p-value of less than 0.05 were considered (FDR 5%). The four terms with the best coverage (proportion of proteins in subnetwork annotated to it) for the largest subnetwork for the clinical dataset are shown in Figure 15.

For comparison, a univariate statistical analysis as well as a regular multivariate PLS analysis was performed on the datasets in this study. The univariate analysis generated no significant proteins at a 5% FDR level, for any of the tested datasets. In a complex study of human cell lines or clinical samples one might expect the effect on proteins to be on several proteins, which might not be picked up by a univariate statistical test. The regular PLS model on the other hand suffers from other problems: How to select the optimal set of variables and how to interpret the results. In the current study, PLS analysis on the clinical dataset generated a list of around 80 significant proteins. The 80 proteins were mapped to the STRING network. The proteins were spread throughout the whole network with no visible clustering and very few links connecting the significant proteins. The proteins from regular PLS analysis were also subject to enrichment analysis, which resulted in no significantly enriched terms at a 5% FDR level. This also indicated that the proteins are not biologically related and represent proteins from very different processes. The subnetwork analysis on the other hand, generates functionally related proteins that are linked to each other in the network. This study is still on-going, the statistical model need to be further validated. The significant subnetworks also need to be validated biologically, both to see that the interactions occur under the conditions studied as well as to verify that the protein subnetworks can be picked up in a larger sample cohort and that they have a biological meaning.

2.4 GENERAL CONCLUSIONS AND FUTURE PERSPECTIVES

Extract more information from available data

The work in **paper I** aimed at showing how additional valuable information can be extracted from existing 2DE data by performing a meta-analysis cross different tumour types. By the workflow for meta-analysis developed in **paper I,** we identified a common protein profile for two malign tumour types, which was not identified when analysing the data sets separately. By combining tumour data sets the identified protein profiles could potentially be used in addressing several clinical questions which are difficult to answer based on analysis of a single study. Common proteins profiles could be changes related to oncogenic processes and could thus be used to better understand tumour biology and address common issues such as malignity, severity, survival and risk of metastasis. The meta-analysis should be used in combination with the separate analysis to distinguish the common protein changes from the unique. The specific proteins that are differently expressed only in a certain tissue type, could help to provide a more certain tumour diagnosis.

A similar meta-analysis approach could have several possible uses. We have in the lab now gathered a large amount of mass spectrometry based proteomics datasets, both on human cancer cell lines and on clinical material from several different tumour types. As the approach is not limited to 2DE per se, the data can be used to draw general conclusions on large scale protein expression in different cell lines and in different clinical material to find unique and common patterns. The approach could also be extended to publically available datasets; the amount of MS based proteomics datasets in public domain resources are increasing [183-185].

Similarly, another use of available in-house datasets is to investigate issues such as technical and biological variation and overlaps between experimental runs in terms of quantities and identifications. By studying the in-house generated datasets we can build up better and more detailed knowledge of our experimental system and better understand the limitations and possibilities. Before planning a new experiment one may consider if the question can be answered by already existing datasets, just by approaching it in a different way.

Another important aspect related to using the full potential of each dataset is the amount of data not explained by current peptide identifications by database search (as much as 70% of generated spectra are not explained). This is partly due to low quality spectra of course, but there is also a big limitation in current database search methods to identify protein isoforms and modified proteins. By developing new methods, like SpliceVista in **paper III**, and re-searching available datasets, we can further explore the data. The analysis in **paper III**, demonstrates detection and splice variant specific quantitation of 939 splice variants on protein level in the cell line dataset, hence improving data output compared to conventional analysis. Nevertheless, the method is limited to identify known splice variants reported in the public repositories. The current version of the EVDB database contains around 80 000 splice variants. Considering that more than 90% of human genes may undergo alternative splicing this probably represents only a small fraction of the total protein splice variants. Furthermore, since splicing variants are temporal and tissue specific the databases are likely to be biased towards well studied tissues and conditions. Despite limitations, by taking into account splice variant information, we have the potential to make new findings, from available data as well as from new data.

Importance of high quality quantitative data

In mass spectrometry based proteomics we are almost always interested in the quantity of the proteins, not only the identity. The goal of proteomics discovery research is often to measure quantitative changes in protein levels between two or more different conditions. It is thus crucial to know what quantitative data we can trust. Bad quality quantitative data will have effects at every step in the following data analysis and it will most likely lead to failure in the validation of the findings. To provide solid scientific data and to save time and money we would like to achieve as correct quantitative information as possible as well as to know the limitations of the quantification. The purpose of **paper II** was to generate a basis for the decision of what protein quantities are reliable and find a way for accurate and precise protein quantification. We developed a methodology that improved protein quantification in shotgun proteomics analysis of complex biological samples and introduced a way to assess quantification for proteins with few peptides.

The result in the current study is a guideline to assess the quality of protein quantities. The methodology we developed in **paper II** is applicable to both other datasets as well to other labelling methods. We suggest including a technical duplicate in each experiment, so that the peptide weights can be calculated based on the errors and variations in the current experiment. By using two iTRAQ channels for the technical replicate samples, six iTRAQ channels could be used for other biological samples. This would mean to sacrifice at most one extra channel for a technical replicate. Replicates in the experiment can also be used for other purposes, for example in the down-stream statistical analysis of finding differently expressed proteins. To be able to interpret the data we need at least duplicate samples of the control samples to account for biological and technical variation. We further suggested that the generated figures and tables could be used as a guideline to set a threshold on protein weights to ascertain reliable protein ratios. This will be especially important for proteins with one or a few peptides for quantification. Generally, proteins with few peptides detected as well as low abundant proteins have the largest relative errors and represent the biggest challenge when it comes to reliable protein quantification. For our own lab, I think it is important that this methodology is included in the standard data analysis pipeline.

Another important aspect when considering the reliability of the protein quantities is the difficulty of current database search methods to infer protein isoforms. In MS based proteomics, the peptides identified for a protein could potentially come from different splice variants with similar sequences. If several splice variants of a gene exist in a sample, the quantitative data for that gene/protein is a mixture of all splice variants. During traditional gene centric or protein centric quantitative data analysis, the quantitative data can thus be wrong. The quantitative data will be very different depending on the abundance of the splice variants in the sample, if there exist one or a few highly abundant splice variants, the quantitative data can be dominated by this. The tool developed in **paper III**, SpliceVista, provides the possibility to do splice variant centric analysis at the protein level. The quantitation of a splice variant is done by quantitation of its splice variant specific peptides, i.e. peptides uniquely mapped to one splice variant. This was in several cases shown to be very different from the gene centric or protein centric analysis. The method allows for identification of splice variant specific quantitative changes related to for example clinical questions.

Considerations in the statistical analysis

A critical problem when working with datasets of 5000-10 000 variables but very few samples, as often the case in discovery proteomics, is the risk of false positives. Further complicating the problem is that the clinical data harbours a large amount of biological variation and that we expect the biological changes of interest to be rather small. All those factors lead to risks of making false discoveries in the statistical analysis. A false discovery will lead to failure during the validation of the finding.

One important step to protect against false discoveries is rigorous validations of the statistical model. For univariate methods, methods for correction for multiple tests are applied. For multivariate methods we have in **paper I** and **IV** used cross-validation to assess the models performance on “new” samples not seen by the model during the optimisation and training phase. This is done to make sure the model and the variables selected are general and not performing well only for a certain subset of the samples. In the best of scenarios the statistical model and selected variables are tested on a completely different set of samples. This is most often not possible due to few samples available. The second best scenario would be to perform two layers of cross-validation, an inner layer to optimise the model and select the variables and an outer layer to test the model and variables on the held out test set. In **paper I** a double cross-validation scheme was used, and the final variables were selected based on stability over cross-validation rounds in the inner layer. In **paper IV** this was not possible since the datasets consisted of too few samples. I believe that the subnetwork method would be improved by including a second layer of cross-validation. Some samples could be set aside and the optimal subnetworks would be tested on those. This would both give a better assessment of the predictive performance of the subnetwork as well as an idea of the stability of the subnetwork optimisation process. The subnetworks found to be significant independent on training set used would represent the most stable ones, most likely to perform well also on a different sample cohort.

A rather recent paper [186] discussed the problem of multiple testing adjustment and FDR in multivariate methods. The use of multiple testing correction is more or less standard in univariate methods, while for multivariate methods no such standard exist. The publication suggests a method to assess FDR for molecular signatures, which should be valid for any multivariate statistical method for variable selection and prediction. The problem is that the multivariate model is often optimized with regards to predictive success, while the study showed that signatures that yield high prediction success may still have a high FDR. If the purpose of using multivariate methods is to select variables (discovery) as in the current study, rather than to predict new samples, one should consider taking this into account and select variables based on FDR and stability rather than prediction success. Traditional multivariate methods for classification based on molecular profiles suffer from the fact that there are often several alternative sets of variables yielding the same predictive outcome, thus making it difficult to select the most biologically relevant set of variables for reproducible performance in a different sample cohort. By basing the selection on FDR and stability rather than prediction success this problem might be reduced. This will hopefully also lead to a higher rate of success during validation of the finding.

One has to keep in mind though, that the statistical validation is not the same as the biological validation. The biological validation is crucial to verify the changes detected by the statistical method, preferably in a larger sample cohort, and also to verify that the finding has a biological meaning. But a sound validation of the statistical model to protect against false discoveries would at least give a chance for the biological validation to be successful. With high rate of false discoveries in the discovery phase one cannot expect the biological validation to be successful since the finding is due to chance rather than a real biological or clinical effect.

A completely different, and perhaps complementary, approach to improve statistical power and reduce risk of false discoveries is to increase the number of samples. The strength of the experimental setup we currently use in the lab is the analytical depth, enabling the identification and quantification of the low abundant proteins. But this is on the cost of the number of samples that can practically be run in an experiment. We thus need alternative more high-throughput approaches to be able to increase sample size. One option would be to analyse fewer of the fractions from IPG-IEF, thereby decreasing runtime on the mass spectrometer. Another option is to do label free analysis, but this method suffers from limitations in identification overlap between runs. We are currently working on methods to improve the stability and overlap of consecutive label free runs.

Network based methods

To move from univariate methods to multivariate statistical methods in the study of cancer proteomics is motivated by that complex biological processes not are driven by individual proteins. One natural continuation in this reasoning is to incorporate pathway and network data in the statistical analysis to let the interaction data steer the multivariate model, as in **paper IV**. The motivation for this would be that the set of proteins acting in a biological process are not random, rather they act in interaction with other proteins in signalling pathways and network modules. The results generated in **paper IV** pointed towards potential advantages of a subnetwork based PLS model as compared to a regular PLS model. The resulting significant proteins were subsets of linked proteins with several biological processes and functions in common. The regular PLS analysis on the other hand, resulted in a set of proteins that all represented very different biological processes and functions. Subnetwork based signatures can thus more readily provide a model of the biological mechanisms studied and be easier to interpret, since they represent functionally coupled proteins rather than a collection of sparse proteins.

Another strength of the network based methods is that by restricting the selection of variables by the network, the risk of false and unstable discoveries might be decreased. The network data provide robustness and prior knowledge that is used to filter the possible variables. This also makes the selection of optimal variables easier, since the problem with multivariate models is often that several models give the same predictive success thus making it difficult to select the optimal set of variables.

The network based models are still a quite new approach though and there is probably room for improvement. One possible additional feature that the subnetwork methods need to account for is the inter-individual patient variability, as suggested in Sandberg et al. [187]. Not all patients can be expected to have changes in all of the proteins in a network. The network method thus needs to allow for a few of the proteins in the network not to be differentially expressed in a subset of the samples [188, 189]. Furthermore, the interaction (edge) between the proteins can also be subject to change and can be as important for the phenotype as the changes of protein (node) levels. Several subnetwork methods accounting for changes in edge activity has been presented over the last few years [140, 175, 190, 191]. An additional weakness of the subnetwork approach is the network itself, which relies on databases of protein interaction data. The number of proteins and links covered in current databases are far from being complete and furthermore they are likely to be error prone and biased towards well studied parts of the interactome [192, 193]. So the network based methods are expected to be more correct as the databases gets larger and higher quality.

From data to results to biological and clinical knowledge

We generate terabytes of data per week with the latest instruments and techniques, but there is still a gap in how to move from data to knowledge in a robust and sound way. To enable the leap from data to biological meaning, system based approaches integrating multiple data types are necessary. The integration of vast experimental data on different levels of the system, DNA/RNA/protein/metabolite, together with knowledge of interactions and pathways are key aspects to be able to create better models of disease and healthy phenotype [135-137, 194-199]. We have seen in numerous publications, and at conferences, that the predictive performance and stability of expression data can be improved by incorporating interactome data, see review in [139]. Studies have shown that even though single genes are not conserved in cancer (and other disease), the pathways are. The information held in affected pathways, could also be used to find new drug targets and alternative treatment regimens based on knowledge from other diseases affecting the same pathway. A challenge is to find clever and systematic ways to use all the prior knowledge available in public databases and to integrate those with molecular expression data. This is of course not an easy task, but it is important to use the vast amount of knowledge already collected. One further challenge of building models of complex systems is to find a balance between accurate models, possible models and useful models, which are often not the same thing.

3 POPULÄRVETENSKAPLIG SAMMANFATTNING

Proteiner är cellens arbetshästar; DNA:t har instruktionen och RNA:t är budbäraren medan proteinerna utför själva arbetet. Proteiner är inblandade i alla biologiska processer i cellen. Proteomet är benämningen av alla proteiner i t.ex. en organism eller en vävnad. Det humana protetomet omfattar alltså alla proteiner som finns i människan. Det humana *genomet* (alla gener) består av ca 20 000 gener, det humana *proteomet* däremot uppskattas till omkring totalt 1 000 000 proteiner. Det beror på att en gen kan ge upphov till flera olika proteiner. Detta sker både genom så kallad alternativ splitsning där en gen klipps ihop till olika proteiner och genom post-translationella modifieringar där ett protein får olika kemiska grupper (modifieringar) på sig. Den enorma komplexiteten gör proteomet svårstuderat. Proteomet är dessutom dynamiskt och i konstant förändring - det varierar med tid, celltyp och omgivande betingelser. Studien av proteomet; dess sammansättning, uttrycksnivå, struktur, funktion och interaktioner; kallas för proteomik. En vanlig metod för storskalig analys av proteiner är att använda masspektrometri. Genom masspektrometri får man information om vilka proteiner ett prov innehåller samt den mängd av proteinet som finns i provet. I en så kallad ”shotgun proteomics” analys klipps proteinerna först ner till peptider (fragment av protein) för att sedan analyseras i masspektrometern.

Detta arbete handlar om cancerproteomik, alltså storskalig studie av förändringar i proteomet som är relaterade till cancer. Syftet med vår forskning är dels att hitta proteiner som kan hjälpa oss bättre förstå cancerbiologin, t.ex. varför det uppstår resistens mot ett läkemedel, dels hitta proteiner som kan användas som biomarkörer. En biomarkör är exempelvis ett protein som kan mätas vid provtagning och vars nivå säger något om patientens sjukdom. T.ex. kan en diagnostisk biomarkör användas för att ställa en diagnos medan en prediktiv biomarkör kan användas för att förutsäga hur en patient kommer att svara på en viss behandling. För trots stora framgångar inom cancerdiagnos och behandling är cancer fortfarande den ledande dödsorsaken i världen. Med hjälp av nya biomarkörer hoppas vi kunna förbättra överlevnaden hos cancerpatienter. En tidigt ställd diagnos är avgörande för att kunna sätta in behandling i ett tidigt skede. Genom att skräddarsy behandlingen efter patienten kan man undvika många biverkningar och slippa förlora värdefull tid på behandlingar med dålig effekt.

Cancer är en komplex sjukdom; flera olika gener är påverkade i utvecklingen från normal cell till cancercell. Den senaste tekniken för att studera genomet och proteomet har ingett mycket hopp om att kunna hitta nya, bättre biomarkörer. Tyvärr har forskningen inom dessa områden ännu inte genererat biomarkörer som kommit till klinisk nytta. Hindren har varit många, speciellt har flera storskaliga studier på senaste tiden kunnat visa på svårigheten med att finna stabila biomarkörer baserat på enstaka gener. När studierna upprepats i en annan provkohort har inte samma gener varit förändrade. Man har dock sett att liknande signalvägar (pathways) är påverkade. Trots att det inte är exakt samma gener så är det alltså i samma del av det cellulära systemet som förändringen skett. Det har således föreslagits att det länge använda uttrycket ”Cancer is a disease of the genes” kanske borde ändras till ”Cancer is a disease of the pathways”.

För att kunna använda proteomikdata på bästa sätt och för att kunna dra giltiga slutsatser från resultaten krävs en rigid kvalitetskontroll och avancerad dataanalys. Först och främst måste vi veta att den kvantitativa datan är pålitlig, alltså att det mått vi har på mängden protein är korrekt. Vi måste även ha rätt statistiska metoder för att analysera datan, för att inte riskera att plocka upp proteiner som felaktigt klassificerats som signifikant ändrade. Vi måste även utveckla metoder för att integrera proteinkvantiteter med annan kunskap om t.ex. interaktioner mellan proteiner (protein-nätverk). Senaste tidens studier har tydligt visat att DNA, RNA och proteiner inte räcker för att fullt förstå sjukdomsmekanismer, eftersom biologiska funktioner är mycket mer komplexa än summan av de individuella komponenterna. För att kunna skapa bättre modeller av sjukdom och hälsa krävs en systembaserad analys, där integration av de olika typerna av data är i fokus.

Det huvudsakliga syftet med mitt doktorandprojekt har varit att utveckla robusta metoder för att välja ut nyckelproteiner, nätverk och signalvägar som är relevanta för kliniska frågeställningar, med proteomikdata som utgångspunkt. Projektet har gått från att fastställa lämpliga gränser för kvantifiering, till förbehandling av data samt metodutveckling för statistisk dataanalays mot målet att generera ett set av nyckelproteiner. Jag har även utvecklat systembaserade metoder för att integrera olika typer av data i syfte att förbättra möjligheten att skapa biologiskt och kliniskt relevant information från proteomikdatan.

I studie I utvecklade vi en meta-analys för att kunna koppla samman befintlig data från humana prostata- och kolontumörer. Syftet var att identifiera proteiner som skiljer mellan normala och tumörprover oberoende av vävnadsursprung. Detta arbetsflöde möjliggjorde upptäckten av en gemensam proteinprofil för två maligna tumörtyper, som inte varit möjligt att fastställa då tumörerna analyserades separat.

Syftet med studie II var att skapa beslutsunderlag för vilka proteinkvantiteter (nivåer) som är tillförlitliga och att hitta ett sätt för noggrann och exakt proteinkvantifiering. Vi utvecklade en metod för förbättrad proteinkvantifiering och introducerade ett sätt att bedöma kvaliteten på kvantifieringen av proteiner. Den experimentella designen och de utvecklade algoritmerna minskade det relativa felet i proteinkvantifiering av komplexa biologiska prover.

I studie III utvecklade vi SpliceVista, ett verktyg för identifiering och visualisering av alternativa splitsningsvarianter i masspektrometridata. Genom att matcha identifierade proteiner mot kända splitsningsvarianter, kan SpliceVista identifiera peptider som är specifika för en viss splitsningsvariant och upptäcka differentiellt uttryckta splitsningsvarianter på proteinnivå.

I studie IV utvecklade vi en nätverksbaserad analys för proteomikdata för att identifiera subnätverk med olika aktivitet mellan grupper av prover. Tanken är att flytta fokus från enskilda proteiner som visar differentiellt uttryck till protein-subnätverk med förändrad aktivitet. Metodiken tillämpades på flera av våra kliniska dataset. Genom att studera proteinuttryck i kontexten av protein-nätverk kunde vi detektera skillnader på en systemnivå och förenkla tolkningen av resultaten från cancerstudier.

4 ACKNOWLEDGEMENTS

There are so many people that have contributed to this thesis work, in one way or the other. I would like to specifically thank some of them:

My main supervisor Janne Lehtiö for being a great group leader, supervisor and research mentor. You have managed to create a research group with high ambitions and goals but still maintained a relaxed and open atmosphere. I have really enjoyed being part of your group!

My co-supervisor Jenny Forshed for always having an answer to my questions, no matter if it is a complicated statistical issue or family related subject. I admire your sound way of looking at research and data analysis and you have taught me so much. You are also a great friend!

All the mass spec group members: AnnSofi, Helena, Hanna, Hillevi, Jessie, Lukas, Maria, Henrik, Rui, Yafeng, Jorrit, Kie, Hassan, Anna, Elena, Davide. Thanks for helpful support, ideas, thoughts and discussions on research, for travelling company to conferences and for nice chats during lunches and "fikas" ☺. Thanks for making this group such a nice place to work!

I would also like to thank SciLifeLab research groups and staff for making it an inspiring and outstanding research environment. Many thanks also to Susanne, Birgitta, Meeri and Rolf and all other past and present members of "KBC".

Tack så mycket också till alla vänner som bidrar med så mycket annat betydelsefullt i livet. Gävletjejerna Stina, Ylva, Sofia och Fia för våra traditionella regelbundna träffar som jag alltid ser fram emot. Vännerna från Uppsala-studierna Josefin och Susanne för samtal om livets viktiga och oviktiga saker. Dagmara, för att du alltid är så positiv trots motgångar i livet, jag är så glad att jag lärt känna dig. Och alla andra goda vänner, även om jag inte nämner er specifikt, ni är toppen!

Tusen tack till min stora fina familj för allt ni gör och är. Mor och far för trygghet, kärlek och support. Mina syskon Jonatan, Hanna, Jakob med sin lilla familj Josefin och Roy och Jonas med sin familj Anna och Hedda. Ni är coola, smarta, roliga och omtänksamma. Ni är så viktiga för mig!

Tack Jakob för det fina omslaget. Du lyckades sammanfatta essensen av mitt arbete i tre bilder, bättre än jag lyckades med på 40 sidor!!

Till sist vill jag tacka de viktigaste personerna i mitt liv, mina tre pojkar Ilan, Lián och Yoél. Det är ni som är mitt liv. Tack Ilan för att du alltid tror på mig och stöttar mig. Jag skulle inte klarat det utan er. *Ani ohevet otchem!*

5 REFERENCES

1. Wasinger, V.C., et al., *Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium.* Electrophoresis, 1995. **16**(7): p. 1090-4.
2. Wilkins, M.R., et al., *Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it.* Biotechnol Genet Eng Rev, 1996. **13**: p. 19-50.
3. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
4. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
5. *A gene-centric human proteome project: HUPO--the Human Proteome organization.* Mol Cell Proteomics, 2010. **9**(2): p. 427-9.
6. Legrain, P., et al., *The human proteome project: current state and future direction.* Mol Cell Proteomics, 2011. **10**(7): p. M111 009993.
7. Gstaiger, M. and R. Aebersold, *Applying mass spectrometry-based proteomics to genetics, genomics and network biology.* Nat Rev Genet, 2009. **10**(9): p. 617-27.
8. Jensen, O.N., *Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry.* Curr Opin Chem Biol, 2004. **8**(1): p. 33-41.
9. Maier, T., M. Guell, and L. Serrano, *Correlation of mRNA and protein in complex biological samples.* FEBS Lett, 2009. **583**(24): p. 3966-73.
10. Schwanhausser, B., et al., *Global quantification of mammalian gene expression control.* Nature, 2011. **473**(7347): p. 337-42.
11. Overington, J.P., B. Al-Lazikani, and A.L. Hopkins, *How many drug targets are there?* Nat Rev Drug Discov, 2006. **5**(12): p. 993-6.
12. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* Cell, 2000. **100**(1): p. 57-70.
13. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
14. Srinivas, P.R., B.S. Kramer, and S. Srivastava, *Trends in biomarker research for cancer detection.* Lancet Oncol, 2001. **2**(11): p. 698-704.
15. Srinivas, P.R., et al., *Proteomics for cancer biomarker discovery.* Clin Chem, 2002. **48**(8): p. 1160-9.
16. Jemal, A., et al., *Global cancer statistics.* CA Cancer J Clin, 2011. **61**(2): p. 69-90.
17. Jimenez, C.R. and M. Slijper, *Current status of cancer proteomics, how far are we from clinical applications?* J Proteomics, 2010. **73**(10): p. 1787-9.
18. Sawyers, C.L., *The cancer biomarker problem.* Nature, 2008. **452**(7187): p. 548-52.
19. Wistuba, II, et al., *Methodological and practical challenges for personalized cancer therapies.* Nat Rev Clin Oncol, 2011. **8**(3): p. 135-41.
20. Poste, G., *Bring on the biomarkers.* Nature, 2011. **469**(7329): p. 156-7.
21. Beretta, L., *Proteomics from the clinical perspective: many hopes and much debate.* Nature Methods, 2007. **4**(10): p. 785-6.
22. Solassol, J., et al., *Clinical proteomics and mass spectrometry profiling for cancer detection.* Expert Rev Proteomics, 2006. **3**(3): p. 311-20.
23. Anderson, N.L. and N.G. Anderson, *The human plasma proteome: history, character, and diagnostic prospects.* Mol Cell Proteomics, 2002. **1**(11): p. 845-67.
24. Schiess, R., B. Wollscheid, and R. Aebersold, *Targeted proteomic strategy for clinical biomarker discovery.* Mol Oncol, 2009. **3**(1): p. 33-44.
25. Ransohoff, D.F., *Bias as a threat to the validity of cancer molecular-marker research.* Nat Rev Cancer, 2005. **5**(2): p. 142-9.
26. Ransohoff, D.F. and M.L. Gourlay, *Sources of bias in specimens for research about molecular markers for cancer.* J Clin Oncol, 2010. **28**(4): p. 698-704.

27. Landegren, U., et al., *Opportunities for sensitive plasma proteome analysis.* Anal Chem, 2012. **84**(4): p. 1824-30.

28. Gorg, A., et al., *The current state of two-dimensional electrophoresis with immobilized pH gradients.* Electrophoresis, 2000. **21**(6): p. 1037-1053.

29. Rabilloud, T., et al., *Two-dimensional gel electrophoresis in proteomics: Past, present and future.* J Proteomics, 2010. **73**(11): p. 2064-77.

30. Rabilloud, T. and C. Lelong, *Two-dimensional gel electrophoresis in proteomics: a tutorial.* J Proteomics, 2011. **74**(10): p. 1829-41.

31. Lilley, K.S., A. Razzaq, and P. Dupree, *Two-dimensional gel electrophoresis: recent advances in sample preparation, detection and quantitation.* Curr Opin Chem Biol, 2002. **6**(1): p. 46-50.

32. Aebersold, R., *Quantitative proteome analysis: methods and applications.* J Infect Dis, 2003. **187 Suppl 2**: p. S315-20.

33. Aebersold, R. and M. Mann, *Mass spectrometry-based proteomics.* Nature, 2003. **422**(6928): p. 198-207.

34. Beck, M., M. Claassen, and R. Aebersold, *Comprehensive proteomics.* Curr Opin Biotechnol, 2011. **22**(1): p. 3-8.

35. Domon, B. and R. Aebersold, *Mass spectrometry and protein analysis.* Science, 2006. **312**(5771): p. 212-7.

36. Mallick, P. and B. Kuster, *Proteomics: a pragmatic perspective.* Nat Biotechnol, 2010. **28**(7): p. 695-709.

37. Mann, M., et al., *The coming age of complete, accurate, and ubiquitous proteomes.* Mol Cell, 2013. **49**(4): p. 583-90.

38. Nilsson, T., et al., *Mass spectrometry in high-throughput proteomics: ready for the big time.* Nature Methods, 2010. **7**(9): p. 681-5.

39. Aebersold, R., *A stress test for mass spectrometry-based proteomics.* Nature Methods, 2009. **6**(6): p. 411-2.

40. Cox, J. and M. Mann, *Quantitative, high-resolution proteomics for data-driven systems biology.* Annu Rev Biochem, 2011. **80**: p. 273-99.

41. Sabido, E., N. Selevsek, and R. Aebersold, *Mass spectrometry-based proteomics for systems biology.* Curr Opin Biotechnol, 2012. **23**(4): p. 591-7.

42. Gevaert, K., et al., *A la carte proteomics with an emphasis on gel-free techniques.* Proteomics, 2007. **7**(16): p. 2698-718.

43. Beck, M., et al., *The quantitative proteome of a human cell line.* Mol Syst Biol, 2011. **7**: p. 549.

44. Geiger, T., et al., *Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins.* Mol Cell Proteomics, 2012. **11**(3): p. M111 014050.

45. Nagaraj, N., et al., *Deep proteome and transcriptome mapping of a human cancer cell line.* Mol Syst Biol, 2011. **7**: p. 548.

46. de Godoy, L.M., et al., *Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast.* Nature, 2008. **455**(7217): p. 1251-4.

47. Gioti, A., et al., *Genomic insights into the atopic eczema-associated skin commensal yeast Malassezia sympodialis.* MBio, 2013. **4**(1): p. e00572-12.

48. Nagaraj, N., et al., *System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap.* Mol Cell Proteomics, 2012. **11**(3): p. M111 013722.

49. Picotti, P., et al., *A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis.* Nature, 2013. **494**(7436): p. 266-70.

50. Hanash, S.M., C.S. Baik, and O. Kallioniemi, *Emerging molecular biomarkers--blood-based strategies to detect and monitor cancer.* Nat Rev Clin Oncol, 2011. **8**(3): p. 142-50.

51. Hanash, S.M., S.J. Pitteri, and V.M. Faca, *Mining the plasma proteome for cancer biomarkers.* Nature, 2008. **452**(7187): p. 571-9.

52. Hawkridge, A.M. and D.C. Muddiman, *Mass spectrometry-based biomarker discovery: toward a global proteome index of individuality.* Annu Rev Anal Chem (Palo Alto Calif), 2009. **2**: p. 265-77.

53. Hood, B.L., N.A. Stewart, and T.P. Conrads, *Development of high-throughput mass spectrometry-based approaches for cancer biomarker discovery and implementation.* Clin Lab Med, 2009. **29**(1): p. 115-38.

54. Kulasingam, V. and E.P. Diamandis, *Strategies for discovering novel cancer biomarkers through utilization of emerging technologies.* Nat Clin Pract Oncol, 2008. **5**(10): p. 588-99.

55. Liotta, L.A. and E.F. Petricoin, *Mass spectrometry-based protein biomarker discovery: solving the remaining challenges to reach the promise of clinical benefit.* Clin Chem, 2010. **56**(10): p. 1641-2.

56. Surinova, S., et al., *On the development of plasma protein biomarkers.* J Proteome Res, 2011. **10**(1): p. 5-16.

57. Taguchi, A. and S.M. Hanash, *Unleashing the power of proteomics to develop blood-based cancer markers.* Clin Chem, 2013. **59**(1): p. 119-26.

58. Wang, P., J.R. Whiteaker, and A.G. Paulovich, *The evolving role of mass spectrometry in cancer biomarker discovery.* Cancer Biol Ther, 2009. **8**(12): p. 1083-94.

59. Borrebaeck, C.A. and C. Wingren, *High-throughput proteomics using antibody microarrays: an update.* Expert Rev Mol Diagn, 2007. **7**(5): p. 673-86.

60. Borrebaeck, C.A. and C. Wingren, *Design of high-density antibody microarrays for disease proteomics: key technological issues.* J Proteomics, 2009. **72**(6): p. 928-35.

61. Olsson, N., et al., *Proteomic analysis and discovery using affinity proteomics and mass spectrometry.* Mol Cell Proteomics, 2011. **10**(10): p. M110 003962.

62. Wingren, C., P. James, and C.A. Borrebaeck, *Strategy for surveying the proteome using affinity proteomics and mass spectrometry.* Proteomics, 2009. **9**(6): p. 1511-7.

63. Stoevesandt, O. and M.J. Taussig, *Affinity proteomics: the role of specific binding reagents in human proteome analysis.* Expert Rev Proteomics, 2012. **9**(4): p. 401-14.

64. Yates, J.R., C.I. Ruse, and A. Nakorchevsky, *Proteomics by mass spectrometry: approaches, advances, and applications.* Annu Rev Biomed Eng, 2009. **11**: p. 49-79.

65. Yates, J.R., *A century of mass spectrometry: from atoms to proteomes.* Nature Methods, 2011. **8**(8): p. 633-637.

66. Franck, J., et al., *MALDI imaging mass spectrometry: state of the art technology in clinical proteomics.* Mol Cell Proteomics, 2009. **8**(9): p. 2023-33.

67. McDonnell, L.A., et al., *Peptide and protein imaging mass spectrometry in cancer research.* J Proteomics, 2010. **73**(10): p. 1921-44.

68. Schwamborn, K., *Imaging mass spectrometry in biomarker discovery and validation.* J Proteomics, 2012. **75**(16): p. 4990-8.

69. Domon, B. and R. Aebersold, *Options and considerations when selecting a quantitative proteomics strategy.* Nat Biotechnol, 2010. **28**(7): p. 710-21.

70. Deutsch, E.W., H. Lam, and R. Aebersold, *Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics.* Physiol Genomics, 2008. **33**(1): p. 18-25.

71. Helsens, K., et al., *Mass spectrometry-driven proteomics: an introduction.* Methods Mol Biol, 2011. **753**: p. 1-27.

72. Righetti, P.G., et al., *Prefractionation techniques in proteome analysis: the mining tools of the third millennium.* Electrophoresis, 2005. **26**(2): p. 297-319.

73. Wu, L. and D.K. Han, *Overcoming the dynamic range problem in mass spectrometry-based shotgun proteomics.* Expert Rev Proteomics, 2006. **3**(6): p. 611-9.

74. Steen, H. and M. Mann, *The ABC's (and XYZ's) of peptide sequencing.* Nature Reviews Molecular Cell Biology, 2004. **5**(9): p. 699-711.

75. Katajamaa, M. and M. Oresic, *Data processing for mass spectrometry-based metabolomics.* J Chromatogr A, 2007. **1158**(1-2): p. 318-28.
76. Martens, L., *Bioinformatics challenges in mass spectrometry-driven proteomics.* Methods Mol Biol, 2011. **753**: p. 359-71.
77. Lemoine, J., et al., *The current status of clinical proteomics and the use of MRM and MRM(3) for biomarker validation.* Expert Rev Mol Diagn, 2012. **12**(4): p. 333-42.
78. Makawita, S. and E.P. Diamandis, *The bottleneck in the cancer biomarker pipeline and protein quantification through mass spectrometry-based approaches: current strategies for candidate verification.* Clin Chem, 2010. **56**(2): p. 212-22.
79. Meng, Z. and T.D. Veenstra, *Targeted mass spectrometry approaches for protein biomarker verification.* J Proteomics, 2011. **74**(12): p. 2650-9.
80. Huttenhain, R., et al., *Perspectives of targeted mass spectrometry for protein biomarker verification.* Curr Opin Chem Biol, 2009. **13**(5-6): p. 518-25.
81. Brody, E.N., et al., *High-content affinity-based proteomics: unlocking protein biomarker discovery.* Expert Rev Mol Diagn, 2010. **10**(8): p. 1013-22.
82. Bantscheff, M., et al., *Quantitative mass spectrometry in proteomics: a critical review.* Anal Bioanal Chem, 2007. **389**(4): p. 1017-31.
83. Ong, S.E. and M. Mann, *Mass spectrometry-based proteomics turns quantitative.* Nat Chem Biol, 2005. **1**(5): p. 252-62.
84. Bachi, A. and T. Bonaldi, *Quantitative proteomics as a new piece of the systems biology puzzle.* J Proteomics, 2008. **71**(3): p. 357-67.
85. Vaudel, M., A. Sickmann, and L. Martens, *Peptide and protein quantification: a map of the minefield.* Proteomics, 2010. **10**(4): p. 650-70.
86. Colaert, N., K. Gevaert, and L. Martens, *RIBAR and xRIBAR: Methods for reproducible relative MS/MS-based label-free protein quantification.* J Proteome Res, 2011. **10**(7): p. 3183-9.
87. Colaert, N., et al., *A comparison of MS2-based label-free quantitative proteomic techniques with regards to accuracy and precision.* Proteomics, 2011. **11**(6): p. 1110-3.
88. Nahnsen, S., et al., *Tools for label-free peptide quantification.* Mol Cell Proteomics, 2013. **12**(3): p. 549-56.
89. Sandin, M., et al., *Generic workflow for quality assessment of quantitative label-free LC-MS analysis.* Proteomics, 2011. **11**(6): p. 1114-24.
90. Gygi, S.P., et al., *Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.* Nat Biotechnol, 1999. **17**(10): p. 994-9.
91. Ross, P.L., et al., *Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents.* Mol Cell Proteomics, 2004. **3**(12): p. 1154-69.
92. Dayon, L., et al., *Relative quantification of proteins in human cerebrospinal fluids by MS/MS using 6-plex isobaric tags.* Anal Chem, 2008. **80**(8): p. 2921-31.
93. Ong, S.E., 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**(5): p. 376-86.
94. Nesvizhskii, A.I., O. Vitek, and R. Aebersold, *Analysis and validation of proteomic data generated by tandem mass spectrometry.* Nature Methods, 2007. **4**(10): p. 787-97.
95. Nesvizhskii, A.I., *Protein identification by tandem mass spectrometry and sequence database searching.* Methods Mol Biol, 2007. **367**: p. 87-119.
96. Jacob, R.J., *Bioinformatics for LC-MS/MS-based proteomics.* Methods Mol Biol, 2010. **658**: p. 61-91.
97. Vaudel, M., A. Sickmann, and L. Martens, *Current methods for global proteome identification.* Expert Rev Proteomics, 2012. **9**(5): p. 519-32.
98. Nesvizhskii, A.I., *A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics.* J Proteomics, 2010. **73**(11): p. 2092-123.

99. Nesvizhskii, A.I. and R. Aebersold, *Interpretation of shotgun proteomic data: the protein inference problem.* Mol Cell Proteomics, 2005. **4**(10): p. 1419-40.

100. Resing, K.A., et al., *Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics.* Anal Chem, 2004. **76**(13): p. 3556-68.

101. Smith, L.M. and N.L. Kelleher, *Proteoform: a single term describing protein complexity.* Nature Methods, 2013. **10**(3): p. 186-7.

102. Nesvizhskii, A.I., et al., *A statistical model for identifying proteins by tandem mass spectrometry.* Anal Chem, 2003. **75**(17): p. 4646-58.

103. Kornblihtt, A.R., et al., *Alternative splicing: a pivotal step between eukaryotic transcription and translation.* Nat Rev Mol Cell Biol, 2013. **14**(3): p. 153-65.

104. Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470-6.

105. Blencowe, B.J., *Alternative splicing: new insights from global analyses.* Cell, 2006. **126**(1): p. 37-47.

106. Garcia-Blanco, M.A., A.P. Baraniak, and E.L. Lasda, *Alternative splicing in disease and therapy.* Nat Biotechnol, 2004. **22**(5): p. 535-46.

107. Omenn, G.S., A.K. Yocum, and R. Menon, *Alternative splice variants, a new class of protein cancer biomarker candidates: findings in pancreatic cancer and breast cancer with systems biology implications.* Dis Markers, 2010. **28**(4): p. 241-51.

108. Muhlberger, I., et al., *Computational analysis workflows for Omics data interpretation.* Methods Mol Biol, 2011. **719**: p. 379-97.

109. Kumar, C. and M. Mann, *Bioinformatics analysis of mass spectrometry-based proteomics data sets.* FEBS Lett, 2009. **583**(11): p. 1703-12.

110. Cappadona, S., et al., *Current challenges in software solutions for mass spectrometry-based quantitative proteomics.* Amino Acids, 2012. **43**(3): p. 1087-108.

111. Kall, L. and O. Vitek, *Computational mass spectrometry-based proteomics.* PLoS Comput Biol, 2011. **7**(12): p. e1002277.

112. Dunkler, D., F. Sanchez-Cabo, and G. Heinze, *Statistical analysis principles for Omics data.* Methods Mol Biol, 2011. **719**: p. 113-31.

113. Smit, S., H.C. Hoefsloot, and A.K. Smilde, *Statistical data processing in clinical proteomics.* J Chromatogr B Analyt Technol Biomed Life Sci, 2008. **866**(1-2): p. 77-88.

114. Noble, W.S., *How does multiple testing correction work?* Nature Biotechnology, 2009. **27**(12): p. 1135-1137.

115. Simes, R.J., *An Improved Bonferroni Procedure for Multiple Tests of Significance.* Biometrika, 1986. **73**(3): p. 751-754.

116. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society Series B-Methodological, 1995. **57**(1): p. 289-300.

117. Hilario, M. and A. Kalousis, *Approaches to dimensionality reduction in proteomic biomarker studies.* Brief Bioinform, 2008. **9**(2): p. 102-18.

118. Jolliffe, I.T., ed. *Principal Component Analysis*. 2nd ed. Springer Series in Statistics. 2002, Springer: New York.

119. Geladi, P. and B.R. Kowalski, *Partial Least-Squares Regression - a Tutorial.* Analytica Chimica Acta, 1986. **185**: p. 1-17.

120. Wold, S., M. Sjostrom, and L. Eriksson, *PLS-regression: a basic tool of chemometrics.* Chemometrics and Intelligent Laboratory Systems, 2001. **58**(2): p. 109-130.

121. Malik, R., et al., *From proteome lists to biological impact--tools and strategies for the analysis of large MS data sets.* Proteomics, 2010. **10**(6): p. 1270-83.

122. Mueller, M., L. Martens, and R. Apweiler, *Annotating the human proteome: beyond establishing a parts list.* Biochim Biophys Acta, 2007. **1774**(2): p. 175-91.

123. Vidal, M., *A unifying view of 21st century systems biology.* FEBS Lett, 2009. **583**(24): p. 3891-4.

124. Sobie, E.A., et al., *Systems biology--biomedical modeling.* Sci Signal, 2011. **4**(190): p. tr2.

125. Bertolaso, M., A. Giuliani, and L. De Gara, *Systems Biology Reveals Biology of Systems.* Complexity, 2010. **16**(6): p. 10-16.

126. Ideker, T., T. Galitski, and L. Hood, *A new approach to decoding life: systems biology.* Annu Rev Genomics Hum Genet, 2001. **2**: p. 343-72.

127. Ein-Dor, L., et al., *Outcome signature genes in breast cancer: is there a unique set?* Bioinformatics, 2005. **21**(2): p. 171-8.

128. Ein-Dor, L., O. Zuk, and E. Domany, *Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer.* Proc Natl Acad Sci U S A, 2006. **103**(15): p. 5923-8.

129. *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061-8.

130. Ding, L., et al., *Somatic mutations affect key pathways in lung adenocarcinoma.* Nature, 2008. **455**(7216): p. 1069-75.

131. Gutierrez, A., et al., *High frequency of PTEN, PI3K, and AKT abnormalities in T-cell acute lymphoblastic leukemia.* Blood, 2009. **114**(3): p. 647-50.

132. Inoki, K., M.N. Corradetti, and K.L. Guan, *Dysregulation of the TSC-mTOR pathway in human disease.* Nat Genet, 2005. **37**(1): p. 19-24.

133. Rual, J.F., et al., *Towards a proteome-scale map of the human protein-protein interaction network.* Nature, 2005. **437**(7062): p. 1173-8.

134. Bader, G.D., M.P. Cary, and C. Sander, *Pathguide: a pathway resource list.* Nucleic Acids Res, 2006. **34**(Database issue): p. D504-6.

135. Barabasi, A.L., N. Gulbahce, and J. Loscalzo, *Network medicine: a network-based approach to human disease.* Nat Rev Genet, 2011. **12**(1): p. 56-68.

136. Auffray, C., Z. Chen, and L. Hood, *Systems medicine: the future of medical genomics and healthcare.* Genome Med, 2009. **1**(1): p. 2.

137. Vidal, M., M.E. Cusick, and A.L. Barabasi, *Interactome networks and human disease.* Cell, 2011. **144**(6): p. 986-98.

138. Chuang, H.Y., et al., *Network-based classification of breast cancer metastasis.* Mol Syst Biol, 2007. **3**: p. 140.

139. Nibbe, R.K., et al., *Protein-protein interaction networks and subnetworks in the biology of disease.* Wiley Interdiscip Rev Syst Biol Med, 2011. **3**(3): p. 357-67.

140. Taylor, I.W., et al., *Dynamic modularity in protein interaction networks predicts breast cancer outcome.* Nat Biotechnol, 2009. **27**(2): p. 199-204.

141. Garrels, J.I., *The Quest System for Quantitative-Analysis of Two-Dimensional Gels.* Journal of Biological Chemistry, 1989. **264**(9): p. 5269-5282.

142. Cargile, B.J. and J.L. Stephenson, Jr., *An alternative to tandem mass spectrometry: isoelectric point and accurate mass for the identification of peptides.* Anal Chem, 2004. **76**(2): p. 267-75.

143. Eriksson, H., et al., *Quantitative membrane proteomics applying narrow range peptide isoelectric focusing for studies of small cell lung cancer resistance mechanisms.* Proteomics, 2008. **8**(15): p. 3008-18.

144. Lengqvist, J., K. Uhlen, and J. Lehtio, *iTRAQ compatibility of peptide immobilized pH gradient isoelectric focusing.* Proteomics, 2007. **7**(11): p. 1746-52.

145. Sevinsky, J.R., et al., *Whole genome searching with shotgun proteomic data: applications for genome annotation.* J Proteome Res, 2008. **7**(1): p. 80-8.

146. Makarov, A., et al., *Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer.* Anal Chem, 2006. **78**(7): p. 2113-20.

147. Perry, R.H., R.G. Cooks, and R.J. Noll, *Orbitrap mass spectrometry: instrumentation, ion motion and applications.* Mass Spectrom Rev, 2008. **27**(6): p. 661-99.

148. Olsen, J.V., et al., *A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed.* Mol Cell Proteomics, 2009. **8**(12): p. 2759-69.

149. Yates, J.R., et al., *Performance of a linear ion trap-orbitrap hybrid for peptide analysis.* Analytical Chemistry, 2006. **78**(2): p. 493-500.

150. Dayon, L., et al., *Combining low- and high-energy tandem mass spectra for optimized peptide quantification with isobaric tags.* J Proteomics, 2010. **73**(4): p. 769-77.

151. Medzihradszky, K.F., et al., *The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer.* Anal Chem, 2000. **72**(3): p. 552-8.

152. Chernushevich, I.V., A.V. Loboda, and B.A. Thomson, *An introduction to quadrupole-time-of-flight mass spectrometry.* J Mass Spectrom, 2001. **36**(8): p. 849-65.

153. Sadygov, R.G., D. Cociorva, and J.R. Yates, 3rd, *Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book.* Nature Methods, 2004. **1**(3): p. 195-202.

154. Flicek, P., et al., *Ensembl 2013.* Nucleic Acids Res, 2013. **41**(Database issue): p. D48-55.

155. Elias, J.E. and S.P. Gygi, *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.* Nature Methods, 2007. **4**(3): p. 207-14.

156. Keller, A., et al., *Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search.* Anal Chem, 2002. **74**(20): p. 5383-92.

157. Storey, J.D. and R. Tibshirani, *Statistical significance for genomewide studies.* Proc Natl Acad Sci U S A, 2003. **100**(16): p. 9440-5.

158. Reiter, L., et al., *Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry.* Mol Cell Proteomics, 2009. **8**(11): p. 2405-17.

159. Power, K.A., et al., *High-throughput proteomics detection of novel splice isoforms in human platelets.* PLoS One, 2009. **4**(3): p. e5001.

160. Kahn, A.B., et al., *SpliceMiner: a high-throughput database implementation of the NCBI Evidence Viewer for microarray splice variant analysis.* BMC Bioinformatics, 2007. **8**: p. 75.

161. Burkhart, J.M., et al., *iTRAQ protein quantification: a quality-controlled workflow.* Proteomics, 2011. **11**(6): p. 1125-34.

162. Forshed, J., et al., *Enhanced Information Output From Shotgun Proteomics Data by Protein Quantification and Peptide Quality Control (PQPQ).* Mol Cell Proteomics, 2011. **10**(10): p. M111 010264.

163. Karp, N.A., et al., *Addressing accuracy and precision issues in iTRAQ quantitation.* Mol Cell Proteomics, 2010. **9**(9): p. 1885-97.

164. Mahoney, D.W., et al., *Relative quantification: characterization of bias, variability and fold changes in mass spectrometry data from iTRAQ-labeled peptides.* J Proteome Res, 2011. **10**(9): p. 4325-33.

165. Lin, W.T., et al., *Multi-Q: a fully automated tool for multiplexed protein quantitation.* J Proteome Res, 2006. **5**(9): p. 2328-38.

166. Gan, C.S., et al., *Technical, experimental, and biological variations in isobaric tags for relative and absolute quantitation (iTRAQ).* J Proteome Res, 2007. **6**(2): p. 821-7.

167. Li, Z., et al., *Systematic Comparison of Label-Free, Metabolic Labeling, and Isobaric Chemical Labeling for Quantitative Proteomics on LTQ Orbitrap Velos.* J Proteome Res, 2012. **11**(3): p. 1582-90.

168. Onsongo, G., et al., *LTQ-iQuant: A freely available software pipeline for automated and accurate protein quantification of isobaric tagged peptide data from LTQ instruments.* Proteomics, 2010. **10**(19): p. 3533-8.

169. Karp, N.A., J.L. Griffin, and K.S. Lilley, *Application of partial least squares discriminant analysis to two-dimensional difference gel studies in expression proteomics.* Proteomics, 2005. **5**(1): p. 81-90.

170. Marengo, E., et al., *Application of partial least squares discriminant analysis and variable selection procedures: a 2D-PAGE proteomic study.* Anal Bioanal Chem, 2008. **390**(5): p. 1327-42.

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

172. Varma, S. and R. Simon, *Bias in error estimation when using cross-validation for model selection.* BMC Bioinformatics, 2006. **7**: p. 91.

173. Fu, W.J., R.J. Carroll, and S. Wang, *Estimating misclassification error with small samples via bootstrap cross-validation.* Bioinformatics, 2005. **21**(9): p. 1979-86.

174. Ideker, T., et al., *Discovering regulatory and signalling circuits in molecular interaction networks.* Bioinformatics, 2002. **18 Suppl 1**: p. S233-40.

175. Kim, Y., et al., *Principal network analysis: identification of subnetworks representing major dynamics using gene expression data.* Bioinformatics, 2011. **27**(3): p. 391-8.

176. Dutkowski, J. and T. Ideker, *Protein networks as logic functions in development and cancer.* PLoS Comput Biol, 2011. **7**(9): p. e1002180.

177. Lavi, O., G. Dror, and R. Shamir, *Network-induced classification kernels for gene expression profile analysis.* J Comput Biol, 2012. **19**(6): p. 694-709.

178. McCormack, T., et al., *Statistical Assessment of Crosstalk Enrichment between Gene Groups in Biological Networks.* PLoS One, 2013. **8**(1): p. e54945.

179. Szklarczyk, D., et al., *The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.* Nucleic Acids Res, 2011. **39**(Database issue): p. D561-8.

180. von Mering, C., et al., *STRING: known and predicted protein-protein associations, integrated and transferred across organisms.* Nucleic Acids Research, 2005. **33**: p. D433-D437.

181. Cline, M.S., et al., *Integration of biological networks and gene expression data using Cytoscape.* Nat Protoc, 2007. **2**(10): p. 2366-82.

182. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.

183. Martens, L., *Resilience in the proteomics data ecosystem: how the field cares for its data.* Proteomics, 2013. **13**(10-11): p. 1548-50.

184. Vizcaino, J.A., J.M. Foster, and L. Martens, *Proteomics data repositories: providing a safe haven for your data and acting as a springboard for further research.* J Proteomics, 2010. **73**(11): p. 2136-46.

185. Gonnelli, G., et al., *Towards a human proteomics atlas.* Anal Bioanal Chem, 2012. **404**(4): p. 1069-77.

186. Nilsson, R., J. Bjorkegren, and J. Tegner, *On reliable discovery of molecular signatures.* BMC Bioinformatics, 2009. **10**: p. 38.

187. Sandberg, A., et al., *Tumor proteomics by multivariate analysis on individual pathway data for characterization of vulvar cancer phenotypes.* Mol Cell Proteomics, 2012. **11**(7): p. M112 016998.

188. Dao, P., et al., *Inferring cancer subnetwork markers using density-constrained biclustering.* Bioinformatics, 2010. **26**(18): p. i625-31.

189. Ulitsky, I., et al., *DEGAS: de novo discovery of dysregulated pathways in human diseases.* PLoS One, 2010. **5**(10): p. e13367.

190. Liu, Y., et al., *Gene interaction enrichment and network analysis to identify dysregulated pathways and their interactions in complex diseases.* BMC Syst Biol, 2012. **6**: p. 65.

191. Ma, H., et al., *COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method.* Bioinformatics, 2011. **27**(9): p. 1290-8.

192. Baker, M., *Proteomics: The interaction map.* Nature, 2012. **484**(7393): p. 271-5.

193. Bonetta, L., *Protein-protein interactions: Interactome under construction.* Nature, 2010. **468**(7325): p. 851-4.

194. Baggs, J.E., M.E. Hughes, and J.B. Hogenesch, *The network as the target.* Wiley Interdisciplinary Reviews-Systems Biology and Medicine, 2010. **2**(2): p. 127-133.

195. Schadt, E.E. and J.L. Bjorkegren, *NEW: network-enabled wisdom in biology, medicine, and health care.* Sci Transl Med, 2012. **4**(115): p. 115rv1.

196. Chen, R. and M. Snyder, *Systems biology: personalized medicine for the future?* Curr Opin Pharmacol, 2012. **12**(5): p. 623-8.

197. Faratian, D., et al., *Cancer systems biology.* Methods Mol Biol, 2010. **662**: p. 245-63.

198. Laubenbacher, R., et al., *A systems biology view of cancer.* Biochim Biophys Acta, 2009. **1796**(2): p. 129-39.

199. Ram, P.T., J. Mendelsohn, and G.B. Mills, *Bioinformatics and systems biology.* Mol Oncol, 2012. **6**(2): p. 147-54.