

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

**CLINICAL PROTEOMICS –  
QUANTITATIVE ANALYSIS AND  
BIOLOGICAL INTERPRETATION**

AnnSofi Sandberg



**Karolinska  
Institutet**

Stockholm 2012

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

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

© AnnSofi Sandberg, 2012

ISBN 978-91-7457-914-7





## ABSTRACT

The main expectations of applying proteomics technologies to clinical questions are the discovery of disease related biomarkers. Despite technological advancement to increase proteome coverage and depth to meet these expectations the number of generated biomarkers for clinical use is small. One of the reasons is that found potential biomarkers often are false discoveries. Small sample sizes, in combination with patient sample heterogeneity increase the risk of false discoveries. To be able to extract relevant biological information from such data, high demands are put on the experimental design and the use of sensitive and quantitatively accurate technologies.

The overall aim of this thesis was to apply quantitative proteomics methods for biomarker discovery in clinical samples. A method for reducing bias by controlling for individual variation in smoking habits is described in **paper I**. The aim of the method was objective assessment of recent smoking in clinical studies on inflammatory responses. In **paper II**, the proteome of alveolar macrophages obtained from smoking subjects with and without the inflammatory lung disease chronic obstructive pulmonary disease (COPD) were quantified by two-dimensional gel-electrophoresis (2-DE). A gender focused analysis showed protein level differences within the female group, with down-regulation of lysosomal pathway and up-regulation of oxidative pathway in COPD patients. **Paper III**, a mass spectrometry based proteomics analysis of tumour samples, contributes to the molecular understanding of vulvar squamous cell carcinoma (VSCC) and we identified a high risk patient subgroup of HPV-negative tumours based on the expression of four proteins, further suggesting that this subgroup is characterized by an altered ubiquitin-proteasome signalling pathway. **Paper III** describes a data analysis workflow for the extraction of biological information from quantitative mass spectrometry based proteomics data. High patient-to-patient tumour proteome variability was addressed by using pathway profiling on individual tumour data, followed by comparison of pathway association ranks in a multivariate analysis. We show that pathway data on individual tumour level can detect subpopulations of patients and identify pathways of specific importance in pre-defined clinical groups by the use of multivariate statistics. In **paper IV**, the potentials and limits of quantitative mass spectrometry on clinical samples was evaluated by defining the quantitative accuracy of isobaric labels and label-free quantification. Quantification by isobaric labels in combination with pI pre-fractionation showed a lower limit of quantification (LOQ) than a label-free analysis without pI pre-fractionation, and 6-plex TMT were more sensitive than 8-plex iTRAQ. Precursor mixing measured by *isolation interference* (MS1 interference) is more linked to the quantitative accuracy of isobaric labels than *reporter ion interference* (MS2 interference). Based on that we could define recommendations for how much isolation interference that can be accepted; in our data <30% isolation interference had little effect the quantitative accuracy.

In conclusion, getting biological knowledge from proteomics studies requires a careful study design, control of possible confounding factors and the use of clinical data to identify disease subtypes. Further, to be able to draw conclusions from the data, the analysis requires accurate quantitative data and robust statistical tools to detect significant protein alterations. Methods around these issues are developed and discussed in this thesis.

## LIST OF PUBLICATIONS

- I. **A. Sandberg**, C. M. Sköld, J. Grunewald, A. Eklund, Å. M. Wheelock. Assessing recent smoking status by measuring exhaled carbon monoxide levels. *PLoS One* 2011;6(12):e28864.
- II. M. Kohler, **A. Sandberg**, S. Kurtovic, A. Thomas, A. Eklund, M. Thevis, C. M. Sköld, Å. M. Wheelock. Gender differences in the bronchoalveolar lavage cell proteome of patients with COPD. *The Journal of Allergy and Clinical Immunology* 2012, Accepted for publication.
- III. **A. Sandberg**, G. Lindell, B. Nordstöm-Källström, R. Branca, K. Gemzell-Danielsson, M. Dahlberg, B. Larson, J. Forshed, J. Lehtiö. Tumor proteomics by multivariate analysis on individual pathway data for characterization of vulvar cancer phenotypes. *Molecular and Cellular Proteomics* 2012, 11.7 M112.016998.
- IV. **A. Sandberg**, R. M. Branca, J. Lehtiö, J. Forshed. Mass spectrometry based protein quantification in complex samples: the impact of labeling and precursor interference. *Manuscript*.

### Publications not included in the thesis

Forshed J, Johansson H.J., Pernemalm M, Branca R. M., **Sandberg A**, Lehtiö J. Enhanced information output from shotgun proteomics data by protein quantification and peptide quality control (PQPQ). *Mol Cell Proteomics*. 2011 Oct;10(10).

Persson G, Sandén T, **Sandberg A**, Widengren J. Fluorescence cross-correlation spectroscopy of a pH-sensitive ratiometric dye for molecular proton exchange studies. *Phys Chem Chem Phys*. 2009 Jun 7;11(21):4410-8.

# CONTENTS

1	Background.....	1
1.1	Clinical proteomics.....	1
1.1.1	Biomarker discovery .....	3
1.1.2	Quantitative proteomics analysis .....	4
1.2	Biological interpretation .....	7
1.3	Cancer .....	8
1.3.1	Protein biomarkers and targeted therapy in cancer .....	9
1.3.2	Human papilloma virus (HPV) induced cancers.....	10
1.4	Smoke induced inflammation and COPD.....	11
2	The present study.....	14
2.1	Aims .....	14
2.2	Material and methods .....	15
2.2.1	Study design .....	15
2.2.2	Pre-fractionation by peptide isoelectric focusing.....	16
2.2.3	Protein quantification by 2D-gel electrophoresis.....	16
2.2.4	Protein quantification by mass spectrometry.....	17
2.2.5	Protein identification and quantification .....	20
2.2.6	Pre-processing of quantitative proteomics data.....	21
2.2.7	Statistical analyses of quantitative data .....	21
2.2.8	Pathway analysis .....	24
2.2.9	Immunohistochemistry.....	24
2.3	Results and discussion.....	25
2.3.1	Paper I.....	25
2.3.2	Paper II.....	27
2.3.3	Paper III .....	29
2.3.4	Paper IV .....	31
2.4	General conclusions and future perspectives.....	33
2.4.1	Methodological conclusions .....	33
2.4.2	Biological conclusions .....	34
2.4.3	Future perspectives.....	34
	Acknowledgements.....	36
3	References.....	37

## LIST OF ABBREVIATIONS

2-DE	Two-dimensional gel electrophoresis
ANOVA	Analysis of variance
BAL	Bronchoalveolar lavage
CID	Collision induced dissociation
CO	Carbon monoxide
COPD	Chronic Obstructive Pulmonary Disease
CV	Cross-validation
EGFR	Epidermal growth factor receptor
ESI	Electrospray ionization
FDR	False discovery rate
FEV <sub>1</sub>	Forced expiratory volume in 1s
FVC	Forced vital capacity
HCD	Higher-energy collisional dissociation
HPV	Human papilloma virus
IEF	Isoelectric focusing
IHC	Immunohistochemistry
IPA	Ingenuity pathway analysis
IPG	Immobilized pH gradient
iTRAQ	Isobaric tags for relative and absolute quantification
LC	Liquid chromatography
LGMN	Legumain
LOQ	Limit of quantification
LTQ	Linear quadropole ion trap
MS	Mass spectrometry
MX1	Myxovirus resistance 1
OPLS	Orthogonal partial least squares analysis
PCA	Principal components analysis
PLS	Partial least squares analysis
PQPQ	Protein quantification by peptide quality control
ROC	Receiver operating curve
SILAC	Stable isotope labeling by amino acids in cell culture
SNP	Single nucleotide polymorphisms
STAT1	Signal transducer and activator of transcription 1
SUS	Shared and unique structures
TMT	Tandem mass tags
UV	Unit variance (scaling)
VIP	Variable influence on projection
VSCC	Vulvar squamous cell carcinoma



# 1 BACKGROUND

## 1.1 CLINICAL PROTEOMICS

Proteins (enzymes, receptors, transcription factors, etc.) are the most immediate effectors of the molecular phenotype and hence investigation of them is critical to understanding the wide spectrum of clinical phenotypes.

Genomic-based studies such as genome-wide profiling for driver mutations, mRNA expression (microarray or RNAseq) profiling, or whole genome association studies investigating associations between single-nucleotide polymorphisms (SNPs) and disease, have defined disease susceptibility genes, and provided important targets for disease classification and biological insight (1). However, in many cases these methods are limited in providing information on clinical prognosis or driving the discovery of new drug targets. The likely reason for this is that DNA and RNA are not on its own the appropriate end-points to study for understanding disease mechanisms or the deregulated molecular pathways underlying them. In addition, cellular pathways are often integrated downstream from genes. Differential expression analysis by mRNA expression profiling gives an approximation to protein abundance alterations and identifies proteins deregulated at the transcriptional level. However, mRNA expression explains the variation in protein expression only to a limited extent (2, 3), and may not capture patterns of posttranslational deregulation. Additionally, proteomic approaches are useful in characterizing the mechanisms of disease, as they can be used to investigate cellular signalling and its pathways.

Clinical proteomics, i.e. the application of proteomics approaches to clinical research, has the potential to successfully move from basic scientific knowledge to clinical applications for the benefit of the patient. The main expectations of applying proteomics technologies to clinical material and questions are regarding the early detection of disease, the prediction of disease development over time and how individual patients will respond to a given treatment, and the identification of novel pharmaceutical targets.

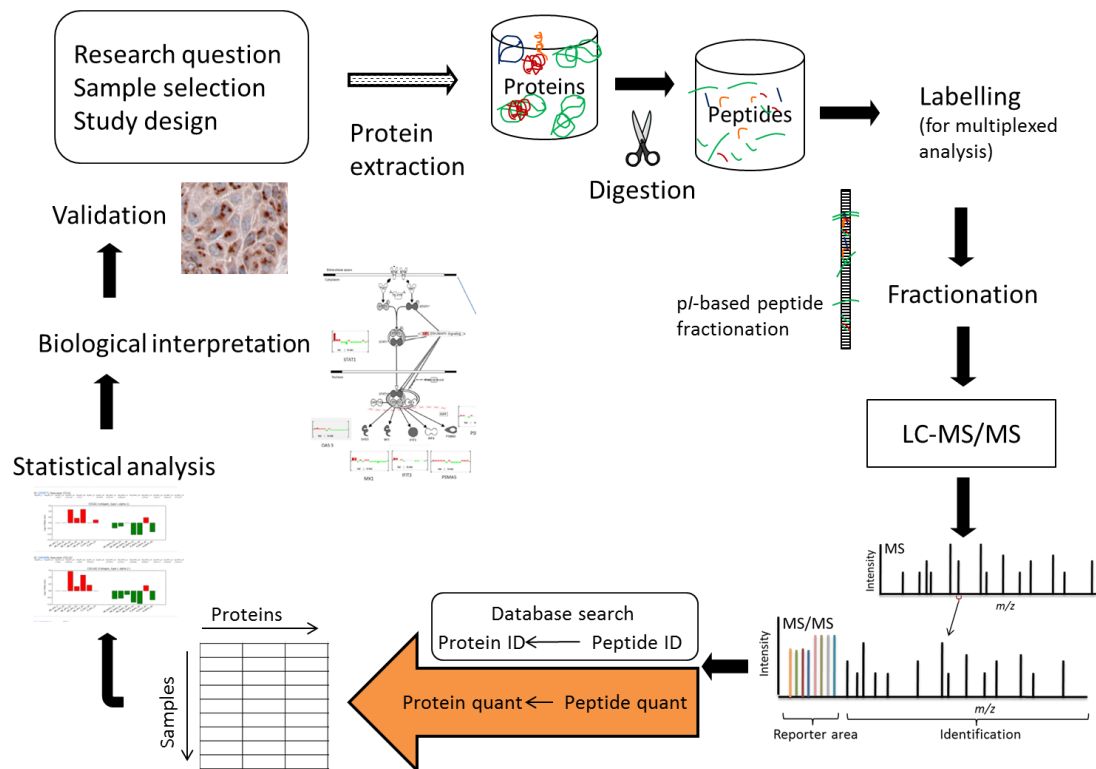
A lot of effort has been put on technological advancement to increase proteome coverage and depth to meet these expectations. With global proteomics technologies, the goal is to analyse the complete set of proteins expressed in a specific organism, tissue or cell at a certain time point of physiological state or experimental condition. Today, a typical 2D-gel electrophoresis (2DE) can visualize approximately 3000 spots (4), although this is without identification. By performing proteomics by liquid-chromatography mass spectrometry (LC-MS) protein profiling, the most advanced labs can detect and identify over 10,000 proteins in a human sample (5, 6). Considering that the number of protein coding genes in the human genome is 20,225 (UniProtKB/SwissProt 2012-10-01), proteomics is approaching proteome coverage in the analysis.

But technological advancement does not solve all the problems. So far, potential biomarkers found by proteomics discovery studies mostly fail at the stage of validation in larger clinical cohorts (7). Moving from potential protein biomarkers in the discovery phase to a biomarker for the patients' benefit is difficult for several reasons. One of the reasons is that the potential biomarkers found often are false discoveries from the data analysis; it is inevitable that technologies like mass spectrometry that survey thousands of different compounds turn up with false positives. Or, proteins can change their abundance correlating with a disease just in the sample cohort. That means candidate biomarkers must be validated, ideally at an early stage to save money and time.

Small sample sizes in combination with patient sample heterogeneity increase the risk of false discoveries. The sample size is limited by the low throughput rate of current state of the art proteomics methods for biomarker discovery (8, 9), which is low due to time-consuming analysis methods required for reaching the low abundant proteins.

To extract relevant biological information from such data, high demands are put on the experimental design and the use of technologies with the sensitivity and quantitative accuracy for low abundance proteins (10), as well as on data analysis. In addition, it is crucial that the samples are of high quality, well characterized and collected according to strict standard operating procedure. Clinical studies are dealing with large inter-individual variation, as well as the variation due to disease. Ways to minimize (unwanted) variability and bias therefore has to be considered in the experimental design to ensure that the clinical characteristics of interest are investigated. In the analysis of tissues and bio-fluids, issues like tissue heterogeneity need to be considered. The experimental design should also consider the availability of samples for validation.

For the data analysis, a biological interpretation of the proteomics results strengthens a statistical analysis. Thus, false negatives can be avoided at an early stage by putting the altered protein levels into a context and by linking related protein level alterations in the biological interpretation. For that, protein quantification is important. In particular, as small protein level changes may have large biological significance it is desirable that the quantitative methods can accurately quantify also small protein changes.



**Figure 1. A quantitative proteomics workflow from research question to biological information.** The workflow includes formulation of clinical research question and sample selection. Indicated in the figure is also a sample pre-fractionation step to reduce complexity. Important further data analysis steps are shown, such as detection of statistically significant protein alterations and software aided validation to the extract relevant biological information. Selected proteins are then validated by an orthogonal method, here exemplified by immunohistochemistry.

### 1.1.1 Biomarker discovery

The aim with biomarker discovery is to find molecular markers (protein or other) correlated with disease or clinical outcome. Potential biomarkers can be identified via different *omics* approaches; in which the whole genome, proteome or metabolome of several (clinical) samples is analysed and evaluated by statistical group comparisons. The molecular quantitative alterations that are determined statistically significant between the defined clinical groups may then be analysed by pathway mapping analysis for biological interpretation. Putative biomarkers are then evaluated on a larger material.

Advances in mass spectrometry (MS), computational data analysis, and the availability of complete sequence databases for many species have done large-scale proteomics analyses possible (11-13). Technical advances has enabled analysis of low abundant proteins and thus, proteomics has become an important field for biomarker discovery (14, 15). It is today possible to quantify low abundant proteins which is the most important in biomarker discovery (16). Although omics-technologies such as proteomics and DNA microarrays have produced an estimated number of 150 000

scientific papers on putative biomarkers, less than 100 have been validated for clinical practice (17).

Biomarkers are wanted for early detection of disease (diagnostic markers), for the prediction of disease development over time (prognostic markers) and to indicate how individual patients will respond to a given treatment (predictive markers). Individualized medicine, where each patient receives tailored therapy, relies on biomarkers (7).

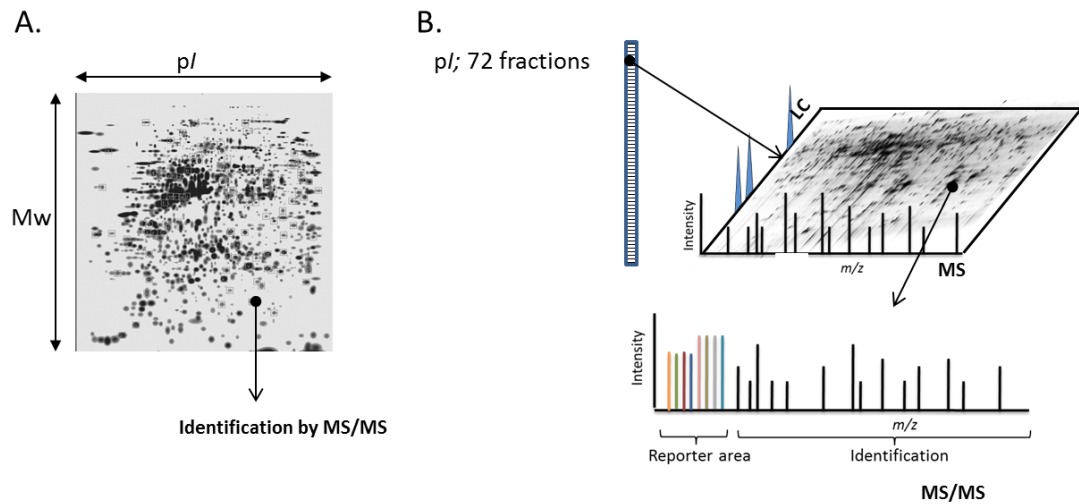
Most clinically used biomarkers are at nanogram [per millilitre plasma] levels or below. One example is prostate specific antigen (PSA) used to diagnose prostate cancer, for which the diagnostic cut-off is 4ng/mL. The normal total protein concentration in plasma is 50-100 mg/mL and total protein concentration ranges 12 orders of magnitude (18). When comparing that with the MS based quantitative proteomics methods, spanning 3-4 orders of magnitude and thus reaching down to low  $\mu\text{g}$  levels (19), it points out one of the difficulties with biomarker discovery. Inter-individual variation is another reason that PSA has limitations in both specificity and sensitivity (20). The overall concentration range in the cell is slightly lower; but still ranges 6-7 orders of magnitude (19, 21). Hence, biomarker discovery studies by proteomics require sub-fractionation of the proteome and sensitive methods to reach the low abundance proteins. However, for the clinical application, i.e. after the discovery phase, antibody based methods are more likely to be used.

### **1.1.2 Quantitative proteomics analysis**

Quantitative data is crucial in biomarker studies. The functionality and significance of the expressed proteins cannot be assessed simply by their qualitative identification. Quantitative data can constitute the absolute number of molecules in a specific sample; for example, when measuring diagnostic biomarkers one may strive to determine their amounts in nanograms per milliliter of blood. Quantitative protein analysis and determining differences in protein levels between two or more sample populations is among the most important tasks in proteomics.

#### *1.1.2.1 Workflow for peptide centric biomarker proteomics*

While a gel based proteomics analysis is performed on intact proteins, the proteome analysis by MS can be performed either on intact proteins (top-down) or on peptides (bottom-up, or shotgun proteomics). Due to the better separation and ionization properties of peptides compared to proteins in both the LC- and the MS-system (22, 23), as well as the fact that peptides can be detected at lower levels compared to proteins (24), peptide-centric proteomics has accomplished the most extensive proteome coverage (25). In this thesis, the quantitative mass spectrometry based methods described are peptide-centric.



**Figure 2. Quantitative proteomics methods.** A. 2-DE workflow: Two-dimensional separation of proteins based on charge ( $pI$ ) and size ( $Mw$ ). Proteins are detected and quantified by post-separation staining or by labelling prior separation with fluorescent dyes. Selected proteins spots are excised, digested into peptides, and identified in a separate step. B. Peptide-centric workflow: Peptide separation by charge ( $pI$ ), hydrophobicity (retention time in liquid chromatography, LC) and by mass ( $m/z$ ) in two steps: first on intact peptides (MS) and then on fragment peptides (MS/MS).

In a peptide-centric approach, a complex protein mixture is enzymatically digested into peptides. The peptides are then often fractionated to reduce sample complexity and increase overall proteome coverage prior to reverse-phase liquid-chromatography (LC) and analysis by mass spectrometry (MS). Sample complexity has an impact on the limit of quantification (LOQ; the limit at which a peptide can be confidently quantified), the dynamic range (the range between the highest signal and the lowest amount of a peptide detected in a single analysis) and the reproducibility of the experiment (26). The LC-MS/MS run is composed of thousands of repeating cycles made up of two basic units: an MS scan that determines the peptide intensity and the mass-to-charge ratios ( $m/z$ ), and an MS/MS scan in which the precursor ion is isolated, fragmented and a spectrum of its fragments is obtained (27). Typically, 5-10 peptides are fragmented for each MS scan. The fragment data from one or up to hundreds of LC-MS/MS runs is then searched in an amino acid sequence database to determine the peptide identity.

The amino acid sequence of the peptides is inferred from the fragment ion spectra. The assignment of fragment ion spectra to peptide sequences is performed by using database search engines such as Mascot or Sequest (28-30) to generate peptide-spectrum matches (PSMs). Identified peptide sequences (PSM's) are then assembled into proteins (31). Protein identifications are thus defined as assemblies of PSMs whose peptide sequences map to the same protein. Protein quantities are then assembled from peptide quantities, often using median values.

Two points in the above-described workflow often create disturbances in biomarker searches. First; neither PSM's nor protein identifications are perfect. To assess the reliability of these identifications, these are therefore controlled for by estimations of false discovery rate (FDR). Second, due to the evolutionary recycling within the genome, identical peptides are occurring at several places in the proteome. This peptide redundancy underlie the protein inference problem (13), which practically means peptide-to-protein mismatching. A consequence of incorrect peptide-protein match is inaccurate quantification on the protein level. There is yet no consensus method for quality control of protein quantification. We use an in-house developed software (32) for the identification and exclusion of peptides judged to be mismatched as determined by their quantitative pattern over several samples. The FDR calculation and protein quantification by peptide quality control used in this study are described in the Methods section.

#### *1.1.2.2 Methods for relative quantification (in MS based proteomics)*

In mass spectrometry, there is a wide range of experimental as well as computational methods for relative quantification, which differ in their accuracy and applicability to various sample types (26, 33-35). MS-based quantification can be performed either using chemical modifications for peptide or protein labelling; or by using label-free approaches on un-labelled peptides.

Another way to classify the quantification methods is based on whether the quantitative data is generated from the precursor ion (MS spectrum) or from the fragments of the precursor ion (MS/MS spectrum). MS based quantification methods are isotopic label methods and label-free quantification by peak intensity or feature detection. MS/MS-based quantification methods are isobaric labelling and label-free quantification by spectral counting.

Label-free quantification methods measure peptide abundance either by comparing peptide spectral counts or by comparing peptide intensities between separate LC-MS/MS runs. In the first case, quantification is done by counting the number of identifications per peptide, which is performed in MS/MS. This is also called spectral counting (36). In the second case, precursor intensities are measured in the MS spectrum by peak area/height (37, 38). With appropriate computational tools, it is then possible to infer the difference in protein abundance between two samples from the precursor intensities (39-41).

The use of labels benefits from multiplexing of the analysis, which reduces instrument time and limits technical variation. The degree of "multiplexing" in a labelling experiment describes how many samples can be differentially labelled, pooled and analysed in one experiment. Labelling of samples with stable isotopes or with isobaric labels creates versions of each peptide that differ in the mass (or fragment mass for isobaric labels) due to isotopic (isobaric tag) composition, but otherwise behave identically during sample preparation, separation and MS analysis. Stable isotope labels can be classified into two broad categories: chemical and metabolic labelling. In

chemical labelling, distinct tags are added to the proteins or the peptides after protein extraction, giving this method the advantage of being applicable to basically all types of protein samples from cells to body fluids (42). In metabolic labelling, heavy isotopes are integrated into the proteome in the process of protein turnover in the living cell (43, 44), with the advantage of minimising experimental bias. In experiments using stable isotope tags, samples are labelled with a “light”, and a “heavy” isotope tag. The relative abundance of each sample can be determined by comparing their relative ion intensities in the mass spectrum. A popular method for metabolic labelling of proteins by isotopic labels is stable isotope labelling by amino acids in cell culture, SILAC (45).

For isobaric labels, the quantification is measured by the intensities of fragment reporter ions from the labels in MS/MS. Isobaric labelling by tandem mass tagging (TMT), and isobaric tags for relative and absolute quantification (iTRAQ), is commonly used in clinical proteomics for peptide and protein quantification. Both TMT and iTRAQ have a high degree of multiplexing which makes them popular. iTRAQ allows 8-fold multiplexing (46), while TMT allows 6-fold multiplexing (47-49). Recent developments has extended TMT to 8-plex (50) by also using isotopic mass shift. An advantage compared to labelling by isotopes is that the isobaric labels will not make the sample more complex than a non-labelled sample.

Both relative and absolute quantitative measurements can be made using stable isotope or isobaric labels. Absolute quantification is often employed as a targeted analysis, i.e. focusing on a single or a few proteins only. A targeted approach is for instance more likely in a validation phase of a study where a few biomarker candidates are to be analysed on a large number of samples.

#### *1.1.2.3 Statistical analysis of the quantitative data*

An in-depth proteomics study helps to identify and quantify as many proteins as possible, but the selection of candidate biomarkers is equally important. Statistical analysis of acquired protein ratios are used to evaluate the significance of the detected differences between clinical groups. In the case of group comparisons and for identifying class discriminating proteins (i.e. a protein or set of proteins that differ between two types of clinical conditions), both Student’s t-test, in which one protein at the time across all samples is evaluated, and supervised multivariate methods which looks at all proteins simultaneously across all samples (such as orthogonal partial least squares analysis, OPLS), can be used to prioritize the combinations of biomarkers that best separates groups of patients. Unsupervised multivariate analysis (such as principal component analysis, PCA) can be used to get an overview of the data and to detect patterns and outliers in the data. The statistical analyses results in a list of proteins that then needs further analysis for obtaining meaningful information.

## **1.2 BIOLOGICAL INTERPRETATION**

To extract meaningful information from the statistical analysis, from which the output often consists of lists of proteins up- or down regulated in different conditions; further

data analysis focused on the biological interpretation of the data is usually needed. Functional analyses by gene ontology (GO) enrichment analysis (51), cellular pathway mapping and network generation (52) can aid in extracting biological information from the quantitative data and put the proteins in a biological context. A biological oriented analysis in terms of pathways or networks is also a validation of the data from the statistical analysis. Software tools for network generation, pathway analysis and GO enrichment analysis are reviewed in (53). Based on the results from a software based validation, a small number of proteins that are most biologically relevant can be singled out and validated experimentally on a larger material by orthogonal methods such as immunohistochemistry or western blot analysis.

Important issues in the pathway analysis are how often the database used to search against is updated and whether it is manually curated or not. Further, pathway mapping and network generation are knowledge based, and not all proteins are represented in the software databases. Another important point is whether the quantitative values (and not just the identities) of the proteins are considered in the analysis.

The biological analysis can be done on sample groups by looking at fold changes and proteins significant for differentiating between defined sample groups as determined by statistical analysis. The biological analysis can also be done on an individual level, with network building of pathway mapping of each individual sample, as described in **paper III**. By using an individual analysis, more proteins can in most cases be used in the analysis since one is not limited to proteins shared by all samples. This increases the statistical strength of the analysis in the pathway mapping. On a biological level, the benefit of an individual pathway analysis is that individual differences are not averaged out.

### 1.3 CANCER

In a multicellular organism, the cell growth needs to be regulated. A cell that does not adapt itself to the needs of the organism but grows autonomously is a cancer cell. The hallmarks of cancer have been summarized to six biological capabilities: sustaining proliferative signalling, evading growth suppression, resisting cell death, replicative immortality, angiogenesis and activating invasion and metastasis (54). Two additional features have been proposed: reprogramming of the energy metabolism and evading immune destruction (55). Further, other tumour infiltrating cells have been emphasized to be of importance for the tumour growth.

The causative agents of cancer is not always known, although many risk factors are known; both lifestyle factors (i.e. smoking) and genetic (i.e. mutant RB1 allele). A driving force for cancer development is genomic instability. Most of the cells' acquired capabilities are consequences of genetic alterations that alter the functions of the protein products of those genes. Despite the diversity in pathogenesis, there are some specific regulatory proteins in the cells that often are disrupted; disruptions of these



lead to acquiring the hallmarks of cancer. Oncogenes (e.g. MYC) are often overactive in cancer, typically due to mutation or overexpression, causing sustained proliferative signalling. In normal cells, the protein products of these genes are involved in cell growth. Tumour suppressor genes (e.g. TP53 coding for the protein p53 and RB1 coding for retinoblastoma protein pRb) are often inactivated in cancer cells. These genes are responsible for the cancer-defence in normal cells; p53 is involved in cell cycle arrest and induction of apoptosis in response to DNA damage, and pRb controls proliferation. Deregulation of p53 and pRb leads to cancer hallmarks such as resistance to cell death and sustained proliferative signalling. The protein p53 is mutated in >50% of all cancers (56).

### **1.3.1 Protein biomarkers and targeted therapy in cancer**

Predicting tumour development and extracting information on tumour-driving molecular changes is one of the biggest challenges in oncology. The hallmarks of cancer are relatively few, but they are the cause of a multitude of phenotypic alterations. Tumour proteomics provides information on phenotype level that combines genetic alterations and environmental effects and therefore gives valuable information guiding tumour characterization. However, comparative proteomic studies on clinical material suffer from large inter-individual variation, and also tumours are heterogeneous (57). This can cause a bias when sampling, as a single biopsy may not be representative for the whole tumour.

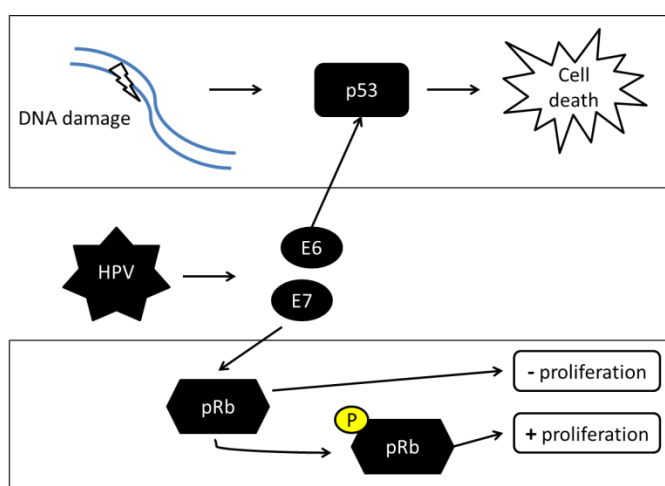
Protein biomarkers currently in clinical use include CA125 for ovarian cancer, CA19-9 for pancreatic cancer, carcinoembryonic antigen for colon cancer and prostate-specific antigen for prostate cancer. However, these all have limited utility for cancer screening (58), largely due to their low sensitivity and specificity in early cancer. Nevertheless, to find biomarkers for early diagnosis, tumour classification and therapy guidance is essential for both individualized medicine (14) and for finding new pharmacological targets. It has been shown for many cancers that the tumours can be divided into subgroups with different properties and prognosis. Thus, the development of new targeted drugs requires the identification of patients with molecularly defined cancers that can be selected for clinical trials to evaluate the new drugs (59).

The hallmarks of cancer have been made targets of cancer therapy. For example, sustaining of proliferative signalling is targeted by inhibitors of epidermal growth factor receptor (EGFR) mediated signalling (60), which can be described as an oncogenic pathway in cancer. Another example is Tamoxifen, a treatment for breast cancer that targets the estrogen receptor (ER). However, clinical responses are often transient and followed by relapse. This could be due to redundant signalling pathways; meaning that the targeted therapy does not completely shut down the hallmark because of parallel pathways that maintain the function. Also, cancer cells within the same tumour can have different cancer driving pathways activated (57).

### 1.3.2 Human papilloma virus (HPV) induced cancers

Although not a major cause of cancer, there are also viral causes of cancer. One such virus is the human papilloma virus (HPV), a sexually transmitted DNA virus that infects squamous epithelium. HPV infection is in most cases transient, but a sub-group of oncogenic HPV strains can cause cancer. HPV is the cause of the majority of cervical carcinomas (99% are HPV positive), and 25 % of head and neck cancers, and play a role also in other malignancies (61). The recently introduced vaccination program against HPV infection is expected to reduce the number of HPV induced cancers in general, although it is developed against cervical cancer and currently restricted to vaccination of girls and young women.

The oncogenic pathways induced by the human papilloma virus (HPV) are well studied; for a review see (62). The oncogenic features of HPV primarily lie in the viral proteins E5, E6 and E7 (Figure 3). E6 and E7 target two of the major cellular tumour suppressor proteins: pRb and p53, which are consequently deregulated in HPV induced cancers (62). E6 and E7 also target the antiviral defences by suppressing interferon (IFN)-mediated immune responses (63). Although much information on HPV induced oncogenic pathways exist on mRNA level, there are few in-depth proteomic studies on clinical material.



**Figure 3. The functional inactivation of tumour suppressor proteins pRb and p53 by the HPV viral proteins E7 and E6.** The protein E7 binds to the retinoblastoma tumour suppressor gene product pRb, with a preference for the active (non-phosphorylated) form of pRb. The E6 proteins associate with the p53 tumour suppressor protein. This interaction promotes the degradation of p53.

#### 1.3.2.1 Vulvar squamous cell carcinoma

Vulvar squamous cell carcinoma (VSCC) is a gynaecological skin tumour. VSCC can be divided into two sub-groups; one HPV positive and one HPV negative (64). The reported proportion of vulvar cancer linked to HPV infection varies widely (65-72), but in Nordic studies 22-52% of VSCC tumours are HPV positive (66, 70, 71, 73). Several studies indicate a positive correlation between HPV related VSCC and favourable

prognosis (70, 73-75). But the link is uncertain as there also are studies showing no prognostic importance of HPV status (68, 76, 77).

Based on clinical and histopathological features the two VSCC subtypes, which are preceded by their own type of pre-malignant lesion, are postulated to develop via separate intracellular signalling pathways (64). The two VSCC types differ in age distribution, with HPV related vulvar carcinoma linked to younger age (approx. 65 yrs.) compared to HPV negative carcinoma (approx. 75 yrs.) (73, 76, 78).

The main treatment of VSCC is surgery, but depending on disease stage, adjuvant radiation and chemotherapy may be given (79). Identification of patients with low risk of relapse could allow for less radical surgery. A report on the first trial of a targeted therapy for VSCC was recently published (80). The evaluated drug was Erlotinib, an EGFR inhibitor. The results, where 27% of the patients showed partial response (which was the best outcome), emphasizes the need for stratification of patients prior to clinical trials.

#### **1.4 SMOKE INDUCED INFLAMMATION AND COPD**

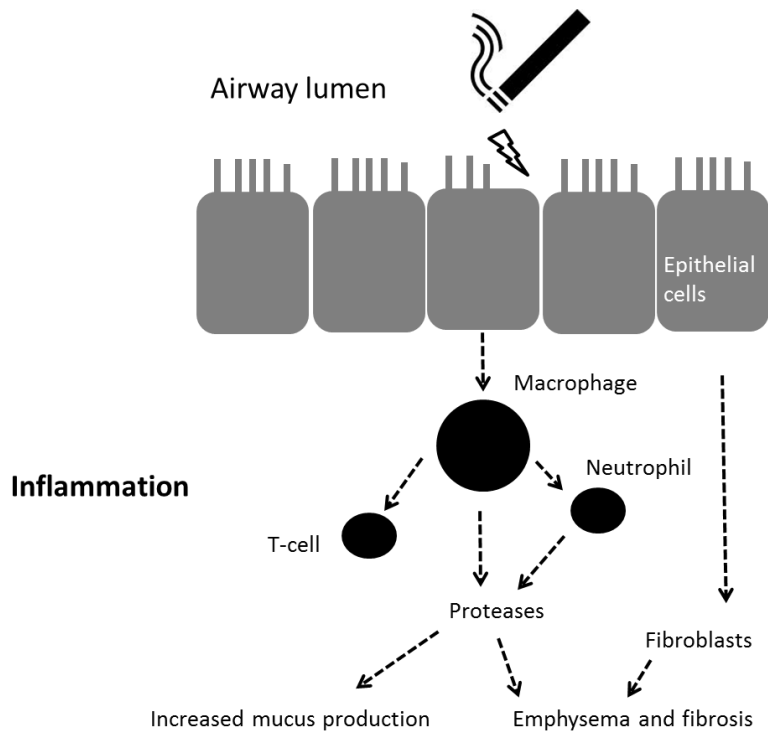
Smoking is a major factor in cardiovascular and chronic lung diseases, and the association of smoking and inflammation is well known (81). The acute effects of smoking have an impact on a number of inflammatory markers in the lung (82, 83). The cigarette smoke induced lung inflammation is a normal response that appears to be modified in patients who develop chronic obstructive pulmonary disease (COPD), which is one of the consequences of smoking and chronic inflammation (84). In these individuals a chronic inflammatory response is induced that may lead to destruction of the lung parenchyma (leading to emphysema) and disruption of repair mechanisms (leading to small airway fibrosis) (85, 86). The consequences are progressive airflow limitation and air trapping. In general, these inflammatory and structural changes of the airways increase with disease severity and persist on smoking cessation.

Clinically, the definition of COPD is:

“a preventable and treatable disease with some significant extrapulmonary effects that may contribute to the severity in individual patients. Its pulmonary component is characterized by airflow limitation that is not fully reversible. The airflow limitation is usually progressive and associated with an abnormal inflammatory response of the lung to noxious particles or gases” (87).

A number of inflammatory cells are involved in the inflammation in COPD (Figure 3), as injury of the airway epithelial cells leads to release of danger-associated molecular patterns and cytokines, which recruits dendritic cells, neutrophils and macrophages to the site (85). Proteolytic enzymes and reactive oxygen species are released which will cause further damage if they are not counterbalanced. The immune response further

involves T-lymphocytes that together with the neutrophils and macrophages release inflammatory mediators and enzymes and interacts with structural cells of the airways (e.g. fibroblasts) and the lung parenchyma (88). Macrophages from COPD patients secrete more inflammatory and elastolytic proteins than those from healthy smokers (89). In contrast to macrophages from never- and healthy smokers, the release of inflammatory mediators is not inhibited by corticosteroids in macrophages from COPD patients (90), rendering the treatment used to maintain other chronic inflammatory airways diseases such as asthma, inefficacious in COPD patients (91).



**Figure 3. Major cells involved in smoke induced inflammation in COPD.** Alveolar macrophages are activated as a consequence of the epithelial damage caused by cigarette smoke. By cytokine release, other immunological cells are activated (neutrophils and T-cells). Neutrophils and macrophages release proteases leading to tissue destruction and perpetuated inflammation. Also fibroblasts are activated by the epithelial damage and show an altered response to the cytokine microenvironment, leading to deregulated repair mechanisms.

It is approximated that up to 50% of all smokers will have developed COPD at the age of 70-80 years (92). This suggests a genetic background, which is strengthened by studies showing an increased risk of airway limitation among siblings to patients with COPD (93), and the identification of a single-nucleotide polymorphism in the gene coding for metalloproteinase 12 as a protective factor (94). Other risk factors for COPD are air pollution (95), tuberculosis (96) and passive smoking (97). Although not studied yet, premature birth may show to be a risk factor for developing COPD later in life, as the pre-and perinatal periods are important for the growth and development of the lung (98).

The prevalence of COPD has historically been higher in men than in women. However, due to the fact that the smoking habits of women have changed, the prevalence is now equal (99). There are also studies indicating that women are more sensitive to smoking than men (100-102), which could be due to that women in general have smaller lungs. In addition, women seem to have more airway symptoms, with thicker small airways, while men have more pronounced emphysema (103).

Diagnosis and classification of COPD is today based on measurements of the lung function by spirometry. For COPD diagnosis, the ratio  $FEV_1/FVC < 0.7$ ; where  $FEV_1$  =forced expiratory volume in 1s, and FVC=forced vital capacity. For classification of disease severity into stages I-IV,  $FEV_1$  as percentage of predicted is used for staging according to the following: stage I is mild disease ( $FEV_1 < 80\%$  of predicted); II moderate ( $FEV_1 50-80\%$ ), III severe ( $FEV_1 30-50\%$ ) and IV (very severe,  $FEV_1 < 30\%$ ).

COPD is presently the fourth leading cause of death world-wide and predicted by WHO to become the third leading cause of death by 2030 (104). Despite being one of the most common chronic diseases in the world; the exact molecular mechanisms of the disease are unknown. The analysis of the mechanisms of the disease is challenged by the presence of multiple phenotypically distinct subgroups within the disease. There is a broad variation in clinical phenotypes of COPD. The pulmonary components of COPD can be divided into two parts: small airway fibrosis which lead to airway obstruction or parenchymal destruction (emphysema) (105). These components are present in varying degree among patients. As indicated by the definition of the disease, COPD also cause systemic inflammation and extrapulmonary manifestations, including ischaemic heart disease, osteoporosis and skeletal muscle wasting (106). Obviously, spirometry does not cover extrapulmonary symptoms or clinical subtypes, which explain why clinical features such as rate of decline in health in patients with COPD do not correlate well with  $FEV_1$  (107). In the new GOLD guidelines for COPD (87) it is now indicated that in the management of the disease also symptoms and history of exacerbations (episodes of disease worsening) should be considered, recognizing the limitations of spirometry alone as diagnostic tool.

There is hence an urgent need for identifying biomarkers for early detection, and for the identification of the different phenotypes of the disease. Besides smoking cessation there is no treatment of COPD that modifies disease progression. Current pharmacotherapy is, with some exceptions, developed for asthma and based on inhaled bronchodilators, and corticosteroids (108). A characterization of the different pathophysiological features behind the subtypes on a molecular level is therefore desirable.

## 2 THE PRESENT STUDY

### 2.1 AIMS

The overall aims of this thesis were to apply quantitative proteomic methods for biomarker discovery in clinical samples. Further; to minimise bias and confounding factors in the sample selection; to develop a data analysis workflow for the extraction of biological information from quantitative mass spectrometry data and to evaluate the potentials and limits of quantitative mass spectrometry based methods.

The specific aims were:

**Paper I:** To use carbon monoxide levels in exhaled air as a tool to discriminate between short term abstinence and continued smoking, and establish a cut-off level for classifying recent smokers from smokers having refrained from smoking >8 hours.

**Paper II:** To identify biomarkers for early detection of chronic obstructive pulmonary disease by investigating the soluble proteome of pulmonary cells by quantitative 2D-gel electrophoresis; focusing on gender-specific protein alterations.

**Paper III:** To identify biomarkers for relapse risk and to investigate the molecular alterations underlying human papilloma virus (HPV) positive and HPV negative vulvar squamous cell carcinoma through in-depth tumour protein profiling by quantitative mass spectrometry.

**Paper IV:** To find the limits and factors affecting the quantitative accuracy and linear range for two principally different and commonly used methods for clinical mass spectrometry-based proteomics: label-free quantification and isobaric labelling.

## 2.2 MATERIAL AND METHODS

This section introduces and describes selected aspects of applied methods. Methodological details are found in each of the **papers I-IV**.

### 2.2.1 Study design

#### 2.2.1.1 Sample selection

**Paper I.** Measurements of carbon monoxide (CO) levels in exhaled air were performed on individuals from 2 study groups: Group 1 consisted of 13 individuals: 6 non-symptomatic current smokers and 7 non-smokers. Group 2 consisted of 86 individuals of which 29 were healthy non-smokers, 38 current smokers with normal lung function and 19 were current smokers with COPD of GOLD stage I and II (mild to moderate disease) (109)). As the study in **paper I** was performed in conjunction to **paper II**, group 2 partially overlap with the study subjects in **paper II**.

**Paper II.** Clinical samples were selected from a cohort of 120 individuals matched in terms of age (45-65 years) and gender. The study subjects consisted of never-smokers with normal lung function, smokers with normal lung function, and COPD patients (current smokers and ex-smokers with mild to moderate disease (GOLD stage I-II,  $FEV_1=50-100\%$  and  $FEV_1/FVC < 0.7$ ). All subjects underwent clinical examination, chest X-ray, CT, and spirometry. COPD patients and healthy smokers were matched in terms of smoking history ( $>10$  pack years (1 pack year corresponds to 20 cigarettes/day for 1 year) and  $>10$  cigarettes/day the past 6 months).

**Paper III.** We selected the clinical samples from a cohort of 37 tumour specimens: 7 human papilloma virus (HPV) positive and 7 HPV negative; with relapse as an independent clinical factor. We could then look at relapse regardless of HPV status. In the pair-wise sample selection we considered 14 clinical variables to match the patients from the two groups.

#### 2.2.1.2 Cell samples

In **paper II**, lung cells were obtained by bronchoalveolar lavage (BAL). During this procedure, a bronchoscope is wedged into a middle-lobe bronchus and a physiological saline solution is instilled and aspirated. The aspirate contains inflammatory cells and airway exudates from the distal parts of the lung. The sampling of cells close to the site of disease is an advantage, as material from the lung otherwise is difficult to access in a non-invasive way. If the fluid recovery is low, the material originates from more proximal parts which alter the BAL cell composition. Therefore, if recovery was  $<40\%$ , the sample was excluded. The cellular fraction of the BAL includes macrophages, lymphocytes and neutrophils. The majority (72-96%) of the BAL cells in healthy individuals are macrophages (110), and the percentage is even higher in smokers;  $>90\%$  (111).

In **paper III** tumour samples were obtained from surgery. Intra-sample heterogeneity is a considerable difficulty when profiling tumour tissue proteome. To avoid intra-

sample heterogeneity caused by infiltrating immune- and stromal cells, we inspected adjacent tissue sections regarding tumour cell percentage during the initial sample selection.

**Paper IV.** The human breast cancer cell line MCF7 was used as representative complex proteome background, into which we spiked 57 standard proteins in amounts spanning 5 orders of magnitude.

#### *2.2.1.3 Sampling of carbon monoxide in exhaled breath*

Carbon monoxide (CO) is a constituent of cigarette smoke that is eliminated almost exclusively via breathing. CO in exhaled air can therefore be used as an indicator of smoking. CO is also produced endogenously during inflammation and can be affected by intake of certain food, but compared to smoking, these effects are relatively small. In **paper I**, CO levels were measured using a portable device, Smokerlyzer Micro EC50 (Bedfont Scientific Ltd, Kent, U.K.). Subjects hold their breath for 20 seconds to allow COHb to form equilibrium with alveolar CO. They then exhale into the mouthpiece of the instrument during which the CO levels are recorded (as ppm). As calculated from triplicate measurements, the CV in our study was <10% for smokers. For non-smokers the CV was 150%, which is explained by the low absolute levels close to the detection limit (0-3 ppm).

#### **2.2.2 Pre-fractionation by peptide isoelectric focusing**

Isoelectric focusing (IEF) involves the separation of proteins or peptides based on their isoelectric point, *pI*. This can be performed by applying the proteins or peptides on an immobilized pH gradient (IPG) gel. An electric current is applied, which causes the protein or peptide to migrate in the IPG gels until reaching the pH where it has no net charge and thus stops migrating (it is focused). Isoelectric focusing is the first dimension of separation in 2-DE. In our mass spectrometry based analytical workflow for protein quantification of complex proteomes, we use high-resolution or narrow range isoelectric focusing of peptides prior to MS analysis to reduce sample complexity and thereby increase proteome coverage (112, 113). By reduction of the sample complexity we are able to detect the low-level proteins (6). After focusing, the peptides are passively eluted into 72 contiguous fractions, using an in-house constructed IPG extractor robotics (GE Healthcare Biosciences AB, prototype instrument). The peptides that focus in the applied pH range represent 96% of the proteome, but have been reduced to one third in number (112). Furthermore, peptide *pI* adds information that aids in identification (114, 115) and can be used to reduce the database search space in the peptide identification (6, 116).

#### **2.2.3 Protein quantification by 2D-gel electrophoresis**

Two-dimensional gel electrophoresis (2-DE) is a proteomics separation method, based on separation of intact proteins with respect to their isoelectric point (*pI*) (by isoelectric focusing) and molecular weight (by SDS-PAGE) (117). The separation of proteins in two dimensions results in a 2D-map of spots where ideally each spot represents an individual protein species. The spots are thereafter visualized by staining and quantified



by specific 2-DE analysis software that matches spots across gels. Selected proteins may be excised and identified using mass spectrometry.

One of the most common methods for relative protein quantification is 2-D difference gel electrophoresis (DIGE) (118). This was the method applied in **paper II**. It allows covalent labelling of samples with three spectrally separated fluorescent dyes, which are co-separated on the same 2-DE gel. This enables analysis of differences in protein abundance between samples, avoiding inter-gel variation by using the same gel. The use of a sample pool as an internal reference standard can facilitate gel-to-gel matching and make it possible to multiplex beyond the number of available labels in larger experimental designs (i.e. compare more than 3 samples) (119). For compatibility with MS-based identification of selected spots, minimal labelling strategies are used (120).

The 2-DE technique has limitations particularly regarding sample throughput; it is labour intensive and time consuming. Further, it is not coupled to identification. Other limitations are difficulties in resolving hydrophobic proteins, proteins with extreme  $pI$  ( $>9.5$ ), large molecular weight proteins ( $>250\,000$  Da) and low abundance proteins. A typical 2-DE can visualize approximately 3000 spots (4), which means that many proteins overlap or co-migrate. Approximately 1000 copies of a protein have to be present in a cell to be detected by 2-DE (121). By using isoelectric focusing with single pH unit strips and very large gels, the detection limit has been reduced to  $>300$  copies/cell with resolution of up to 6000 protein spots (122, 123).

## 2.2.4 Protein quantification by mass spectrometry

### 2.2.4.1 *The mass spectrometer*

Mass spectrometry separates gas phase ions in vacuum based on their mass-to-charge ratio ( $m/z$ ). Basically, the mass spectrometer consists of three parts: an ion source, a mass analyser and a detector. In the ion source, the peptides (or other analytes) become ionized, which is a requirement for their subsequent separation in the mass analyser. Electrospray ionization (ESI) is a frequently used ion source because it is a soft ionization method (i.e. the ionization process leaves the molecular ions intact) and because of the added benefit of direct coupling to reversed-phase liquid chromatography (RPLC). RPLC is a method based on hydrophobicity ( $C_{18}$  columns) and is used for sample clean-up and separation of peptides prior to MS analysis. Nano-ESI, the low flow rate (e.g. 400 nl/min) version of ESI was the ionization technique used in this thesis (**papers III and IV**).

In the mass analyser, the ions are separated based on their  $m/z$ . There are several types of mass analysers, all which have their pros and cons. The instrument used for the mass spectrometry analyses in this thesis (**paper III and IV**) was an **LTQ-Orbitrap**. The LTQ-Orbitraps are hybrid instruments (124, 125), having two kinds of mass analysers. The LTQ (linear quadrupole ion trap) traps the ions in “packets”, and is a modification of a quadrupole ion trap instruments, with improved sensitivity in the low mass range. Other strengths of the ion trap are speed and large trapping capacity. In the ion trap, tandem mass spectra (MS/MS) for the purpose of peptide identification are generated

by collision induced dissociation (CID). In addition to the standard CID, the LTQ-Orbitrap is capable of a special fragmentation method (higher-energy collisional dissociation, HCD) that makes it suitable for the use of isobaric tags (126-132). The Orbitrap mass analyser measures mass by the way the ions oscillate in an electrostatic field; the frequency of the oscillations is converted to  $m/z$  by Fourier transformation. The strengths of the Orbitrap are high mass accuracy, high resolving power, high sensitivity and wide dynamic range.

The ion trap and the Orbitrap use different kinds of detection. In the ion trap, after separation in the mass analyser by their  $m/z$ , the ions hit the detector (a type of electron multiplier, that amplifies the signal of each ion hit), which registers the number of hits at any given  $m/z$  value. In the Orbitrap, the detector consists of two metal electrodes that register the oscillation of the ions as they pass by one or the other, thereby producing a sine wave type of signal. This transient signal is then reverse Fourier transformed into the usual mass spectrum with  $m/z$  on the x-axis and ion count on the y-axis.

#### 2.2.4.2 *The mass spectrum*

The mass spectrum is an abundance plot of mass separated ions. The y-axis shows ion counts (intensity). The x-axis shows mass-to-charge ratio ( $m/z$ ), i.e. mass units divided by the charge of the ion, which when using ESI can carry several charges. In acidic solutions amine groups will be protonated and positively charged. Because trypsin cleaves after arginine and lysines residues, tryptic peptides will be able to carry at least two positive charges, i.e. both at the peptide N-terminus and the arginine or lysine side-chain. An LC-MS experiment also has a time dimension, as mass spectra are recorded continuously during LC-separation. Each time point in the LC chromatogram corresponds to a single mass spectrum recorded at that time.

#### 2.2.4.3 *Quantification in MS and MS/MS*

For a peptide, the quantitative response can be obtained by determining the area under curve for the LC-MS peak, or by measuring peak height. As a rule of thumb about 15 data points (individual spectra) should be recorded across the peak for sufficient quantitative accuracy. Thus it may be necessary to adjust the scan speed (the time used to record a single mass spectrum) depending on the speed of the chromatographic separation. Making several measurements will improve quantitative accuracy as well as mass accuracy. If the sample is complex there is a risk that two co-eluting peptides will have the same mass. Overlapping peaks shift the observed mass value and cause inaccurate quantification. For quantification very low intensity peaks can be difficult to measure. Also very high intensity peaks can be problematic because at some point the detector will reach saturation.

A quantitative measure can also be obtained from the MS/MS fragment spectra. This can be done by counting the number of MS/MS fragmentations for a peptide (spectral counting) or by relative quantification of reporter ions generated from the fragmentation of isobaric labels as described. The advantage with MS/MS

quantification is that the MS/MS spectra contain less noise, which leads to an increased signal-to-noise ratio. Another advantage is that both peptide identification and quantification can be performed on the same tandem mass spectrum, simplifying optimization. Quantification may however be skewed due to co-isolation of other ions which are also fragmented, i.e. precursor mixing. In addition, LC-MS instruments require time to generate an MS/MS spectrum. For samples of high complexity, peptides eluting at the same time as another peptide is being fragmented will be missed. So even if the same sample is run twice, there is no guarantee that all of the same peptides will be fragmented in the second run, which may lead to poor overlap between LC-MS/MS runs.

#### 2.2.4.4 *Quantification by isobaric labelling*

In **papers III** and **IV**, MS-based quantification was performed by isobaric labelling. Isobaric labels consist of three parts: a *linker* (an N-hydroxy succinimide ester group that makes them reactive towards N-terminal amines and the amine group of lysine residues), a *reporter* (with a mass unique for each tag) and a *balance* group (adjusted so that the net mass is identical among the tags). Peptides labelled with isobaric tags are thus equal in mass and will be observed as a single peak in MS. When fragmented, the reporter group of the tag is released and the corresponding reporter ions are observed in the low mass region ( $m/z$  range 113–121 for iTRAQ, and 126–131 for TMT). Quantification is then based on the relative intensities of those reporter ions. Thus, to be quantified, a peptide has to be fragmented to generate MS/MS data.

Co-selection of precursors leads to unidentified fragments in the MS/MS spectra, and thereby lowers the number of identifications (133, 134). It also has a negative effect on quantification, as the reporter ions from co-selected precursor ions superimpose on the “true” reporter ions. Because most of the proteins in a biological sample are unregulated, the co-isolated peptides often create reporter tags with equal relative intensity. Consequently, precursor co-isolation has in several studies been shown to cause systematic underestimation of ratios (131, 135, 136). In an experiment where the aim is to detect proteins that exhibit quantitative changes, ratio compression can lead to false negatives. The problem with poor accuracy linked to precursor mixing was recently discussed in a review (137). The frequency of co-fragmentation is increased by high sample complexity, and consequently reduced by sample pre-fractionation (138).

Other considerations in chemical labelling include side-reactions of the label. In the original publication (139), the authors stated low degrees (<3 %) of tyrosine derivatisation and of un-reacted N-terminal and lysine amines.

#### 2.2.4.5 *Label-free peptide quantification*

Measuring *precursor ion intensities* in the first MS dimension (37, 38) is considered the most accurate quantification of low abundant proteins (140). These are the proteins of interest in most biomarker discovery, and this was the label-free quantification strategy that we evaluated in **paper IV**. *Precursor ion intensities* are calculated from either peak height or peak area of ion count for peptide precursor peaks measured during LC

elution. In the software that we used in the label-free quantification (MaxQuant) in **paper IV**, peptide abundance is estimated from peak area.

Comparisons between LC-MS/MS analyses are performed by matching precursor peaks defined by LC retention time and  $m/z$  between samples. These areas/peaks defined by chromatographic retention time and  $m/z$  are also called MS1 features. The MS1 features are matched across different LC-MS/MS analyses for relative quantification. The matching puts high demands on reproducibility and good alignment-methods to be able to match features. Shifts due to technical variance occur in  $m/z$ , LC retention time and intensity. By calibration of the mass spectrometer,  $m/z$  shifts are typically very small, while drifts in the retention time are larger and may require more sophisticated alignment, especially in larger studies. The latter is crucial particularly in clinical studies where the individual variation makes it necessary to analyse samples from many individuals to be able to pick up disease specific differences in protein amounts between sample groups. Intensity normalization to make abundances comparable across samples is often done by global normalization to equal mean or median ion counts between samples (141). Comparisons have shown that linear normalization regression methods work well for LC-MS data (142, 143), and may be improved by including run order in the normalization (142).

## **2.2.5 Protein identification and quantification**

### *2.2.5.1 Estimating false discovery rates (FDR)*

The peptide spectrum-matches (PSM's) from database matching of spectra and peptide sequences are associated with false positives, particularly when the database is large. The fraction of false positives among the true positives (false discovery rate, FDR) is usually estimated using a target-decoy database containing the real database plus a database of equal size to which no matches are expected (for example the reverse sequences of the "target" database). A cut-off level giving an acceptable FDR is then set; common for biological samples are setting cut-off at 5% FDR.

There is a distinction between the FDR of PSM's and of protein identifications. Depending on whether the statistical analysis is performed on protein or peptide level, peptide or protein FDR is preferred. While setting protein-identification FDR is critical in applications such as biomarker discovery or proteome profiling, setting peptide-identification FDR is important for label-free quantification (144). The biologically relevant outcomes of a peptide-centric proteomics approach are on the protein level. To estimate the false discovery rate of the protein identifications in **Paper III**, we used a specific software for estimating the false discovery rate on protein level, MAYU (145). In **paper IV**, we calculated the peptide and protein FDR in MaxQuant (45).

### *2.2.5.2 Protein quantification by peptide quality control*

An in-lab developed algorithm, PQPQ (32) was used to improve the quantitative accuracy by the identification of protein isoforms. PQPQ checks all the peptides matched to a protein by analysing the quantitative pattern over samples. Based on the assumption that the quantitative pattern of peptides from the same protein should

correlate, PQQ identifies outlier peptides and clusters of peptides with differing patterns over samples.

### 2.2.6 Pre-processing of quantitative proteomics data

The output data from the MS instrument is already pre-processed by the instrument software to a certain extent. The label-free data is acquired in profile mode. The same can be said for the gel based data, which is based on spot volume, where the imaging software removes artefacts. The MS/MS data used in the isobaric quantification is acquired in centroid mode, which means that the data is reduced to peak height.

**Mean centering** of the data (**paper II, III**), is done to shift the quantitative data towards the mean in multivariate analysis. Mean centering is done by taking the average value of each variable (protein) and subtract the obtained value from each data point.

**Scaling.** Unit variance (UV) scaling of the quantitative data (**paper II and III**) is done to make each variable equally important (i.e. a protein level alteration that is small becomes as relevant as a protein level alteration that is large). UV scaling is based on the standard deviation of each variable (protein).

**Normalisation** of the samples can be done based on different assumptions depending on study and sample. In **paper IV** normalisation of label-free quantification data was performed against the total intensity of the LC-MS run, assuming equal peptide amount. By normalization we corrected for differences in overall intensities due to instrument drift during the analyses. In **paper I** normalisation of CO values was performed assuming equal reduction of CO elimination rates during the night in all individuals. By normalization we corrected for differences in at which time of the day sampling was performed on the two study groups

**Log-transformation** of quantitative data can be done to make the data normally distributed. It was for example done in **paper IV** to plot the LOQ curves. A risk with transformation is that noise may be enhanced.

### 2.2.7 Statistical analyses of quantitative data

#### 2.2.7.1 *Students t-tests and related nonparametric tests*

T-tests and corresponding nonparametric tests compare two sets of measured data. For comparing three or more groups, ANOVA or related nonparametric tests are used.

#### **Parametric tests**

Parametric tests (**t-tests** and **ANOVA**) assume that the data follow normal distribution. If the data is not normally distributed, log-transformation of the data can be used to make it normal distributed; which was performed in **papers II and IV**.

**Students t-test:** A t-test show the probability that a test is true. This also includes a small probability that the test is false. If many significance tests are done, as in a

proteomics study (one per protein), a correction for multiple testing is required to control for false positives. This is often done by estimations of false discovery rate (FDR). There are tools for t-tests that correct for multiple testing such as significance analysis of microarray (SAM) (146, 147) used in **paper III**. SAM performs t-test with permutation based correction for multiple testing. Another tool is Mass Conductor: <http://translationalmedicine.stanford.edu/Mass-Conductor>; a web tool for calculating t-test and local FDR. The latter method for estimating FDR is preferred when if the sample groups are small, i.e. <5 individuals in each group.

**ANOVA:** is a statistical test of whether or not the means of several groups are all equal, and therefore generalizes t-test to more than two groups. ANOVA was used to test if the separation of clinical groups were significant, based on the cross-validation of the OPLS-model in **paper III**.

#### **Non-parametric tests**

Non-parametric tests do not assume normal distribution. In this approach, values are ranked from low to high, and the analyses are based on the distribution of ranks. Thus, the nonparametric test only looks at rank and makes no difference whether the values higher or a lot higher.

**Mann-Whitney test:** is a nonparametric test to compare two groups (that are not including paired data). It was used in **paper I**.

**Kruskal-Wallis test:** a nonparametric test that compares the means of three or more unmatched groups. The Kruskal-Wallis test can only tell you that there is an overall significant difference. The Dunn's post test then compares each pair of groups and tells which of these pairwise differences are significant. It was used in **paper I**.

**Spearman correlation;** used in **paper I**, is a non-parametric test to calculate correlation coefficients. Pearson is the corresponding parametric correlation.

#### *2.2.7.2 Regression analysis*

**Linear regression:** The goal of linear regression is to adjust the values of slope and intercept to find the line that best predicts Y from X.

**Non-linear regression:** The goal of nonlinear regression is to fit a model to XY data. The model is expressed as an equation that defines Y as a function of X and one or more parameters.

#### *2.2.7.3 Multivariate statistical analysis*

**Principal component analysis (PCA).** Principal component analysis is an unsupervised statistical method used to describe the variation in a data set, regardless of the sample type. PCA is used to get an overview of the data, to detect clustering of the data and to identify outliers. In a PCA the data matrix, which consists of observations (samples) in rows and variables (proteins) in columns, is plotted along orthogonal

vectors (principal components). The first component describes the largest variation in the data; and is a line that approximates the data by linear regression, fitted using the least squares approach. The second component is orthogonal to the first and describes the second largest variation in the data and so on. The number of principal components describing the data depends on the variation in the data set. Scores describe the observations, and loadings are used to describe the relation among the variables in the data set. The further away from the origin a score or a loading is; the stronger is its impact on the model. The loadings-plot show which x-variables that contribute to the variation among the observations. PCA was performed in **papers II and III**.

**Orthogonal partial least squares methods (OPLS).** The OPLS is a supervised method that describes the variation across samples; with X as the variables (proteins) and Y as the response (i.e. relapse or non-relapse). The X variables that are the most important for predicting Y, i.e. the variables that co-varies with the response, are modelled in the first OPLS component. The variance in X that is orthogonal to Y (uncorrelated variation or noise) is modelled separately in orthogonal PLS components. OPLS was performed in **papers II and III**.

**Cross validation.** Cross validation was used in **paper I** for validating the CO decline models, and in **paper III** for validation of the OPLS models. Cross-validation is a way of testing a model with a test set and a training set despite of limited sample number. In cross validation, one part of the data is held out from the analysis while a model is built based on the rest of the data. The generated model is then used to predict the held out data and the predictive ability is calculated. This is repeated until all samples have been predicted once. ANOVA of the cross validated models is then performed to obtain a p-value indicating the probability that the model is the result of chance only.

**Variable influence on projection (VIP) –plot.** The VIP-plot shows the influence of each X-variable on the (OPLS) model. The amount of explained Y-variable is taken into account. The VIP plot was used to select important variable to optimize the OPLS models in **papers II and III**.

**Analysis of shared and unique structures (SUS-plot).** The SUS-plot is a visualization tool used to identify shared and unique loadings between two different OPLS models. It was used in **paper II**. Variables (proteins) on the diagonal show that the same proteins that are affecting both models. Loadings that do not correlate are unique for the respective OPLS model.

**PLS inner relation** was used in **paper II** to evaluate if Y (clinical parameters on BAL cell content) co-varied with the X variables (proteins).

#### *2.2.7.4 Receiver Operator characteristics (ROC) analysis*

The receiver-operator characteristic (ROC) curves are used to visualize the trade-off between sensitivity and specificity for different cut-off values for discriminating

between two groups. In **Paper I**, ROC curves were used to select cut-off for differentiating recent smokers from those that had refrained from smoking.

The ROC curves plots sensitivity vs. 1-specificity. Sensitivity is defined as the fraction of true positives, and the specificity as the fraction of true negatives. The area under the ROC curve quantifies the overall ability of the test. A test that is no better than flipping a coin has an area of 0.5. A perfect test has an area of 1.00.

### 2.2.8 Pathway analysis

Matching to canonical pathways was performed in **paper III**. By matching of the experimental protein dataset to canonical pathways, the proteins are matched and displayed within well-established signalling or metabolic pathways. We used a web based software from Ingenuity Systems (Ingenuity Pathway Analysis, IPA, [www.ingenuity.com](http://www.ingenuity.com)). In most pathway analysis software, the experimental data (protein ID's) are mapped against the database of well known (canonical) pathways. The canonical pathways most enriched among the proteins are identified, and then displayed in order of statistical significance. In IPA, the degree of matching is ranked by  $-\log(p)$ , where  $p$  is a measure of the probability that the pathway is associated with the dataset by random chance (Fisher's exact test, right-tailed). The higher the  $-\log(p)$ , the stronger is the canonical pathway-dataset matching. The statistical significance ( $p$ -values) of the pathways returned is determined by calculating the extent to which the pathways associated with your dataset deviate from what was expected by chance alone.

It can be noted that the pathway naming is not ontology driven, which means that there can be different pathway names for different softwares.

### 2.2.9 Immunohistochemistry

In **paper III**, we used immunohistochemistry (IHC) staining as an orthogonal method to validate the mass spectrometry data. We selected our antibodies from the Human Protein Atlas, <http://www.proteinatlas.org/> from where we also selected the positive and negative controls (148). IHC links the quantitative protein information to cell-type and sub-cellular localisation. This information is lost in the mass spectrometry analysis unless preceded by cell sorting. IHC staining depends on multiple factors, such as antibody specificity, antigen retrieval, fixation of tissue etc. IHC is a semi-quantitative method. Therefore percentage of staining is preferred to staining intensity, which very much depends on the experimental conditions.



## 2.3 RESULTS AND DISCUSSION

### 2.3.1 Paper I

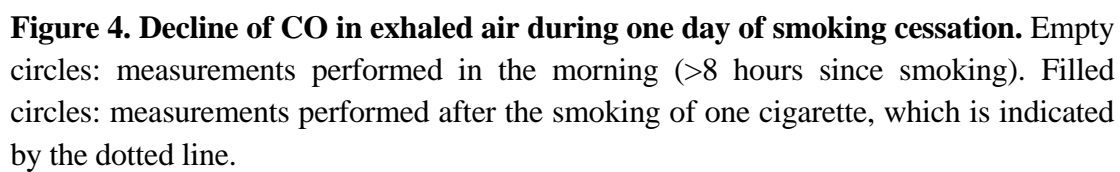
#### *Assessing recent smoking status by measuring exhaled carbon monoxide levels*

An objective way to classify smokers from non-smokers is measuring CO, a constituent of cigarette smoke, in exhaled air (149). With this study our aim was to evaluate whether carbon monoxide (CO) levels in breath could be used to discriminate between short term abstinence and continued smoking. The goal was to establish an objective measure of the study subjects' recent smoking status, to be used in clinical studies on chronic inflammatory disorders.

We performed an initial study on 6 smoking individuals, and monitored CO decline by hourly measurements during one day of smoking abstinence (Figure 4). The decay could be described as a mono-exponential decay ( $r^2 = 0.7$ ) with a half-life of 4.5 hours, in agreement with the literature (150). We also measured CO decline on a second study group consisting of 38 smokers with normal lung function, and 19 smokers with COPD. As CO is produced also during oxidative stress and inflammation (149), we investigated the potential impact of endogenously produced CO from the COPD-related inflammation. However, no significant difference in CO was detected when comparing smokers with normal lung function and smokers with COPD. This is in agreement with that increased levels of CO in exhaled air are associated with exacerbations of the diseases (151). In our study, exacerbations within three months prior to sampling were an exclusion factor.

In the second study group, CO was measured at three separate visits to the clinic. Hence, this data was discontinuous. An initial model of decline gave longer CO half-life (7-9 hrs) compared to the first study. Based on literature showing that CO declines faster during the day (152), we chose to use a correction factor of 1.33 for measurements performed in the afternoon. This correction factor was based on the measurements performed on the first study group. After normalisation, the half-lives were comparable between the two groups. We tested the upper 95% prediction limit of the decline model at 8 hours since smoking, corresponding to 12 ppm CO, as a cut-off by receiver operator curve (ROC) analysis. The 12 ppm cut-off gave a model of high specificity (94%) and sensitivity (90%).

Objective measures of smoking habits are desirable in clinical studies on chronic inflammation as smoking has acute effects on the immune system (153). Due to individual differences in smoking habits it is difficult to set generic cut-offs, however individual cut-offs are not practical in clinical settings. Further evaluation of the proposed cut-off should include additional measurements performed in the critical time range 3-7 hours since smoking, for which very few measurements were performed.



### 2.3.2 Paper II

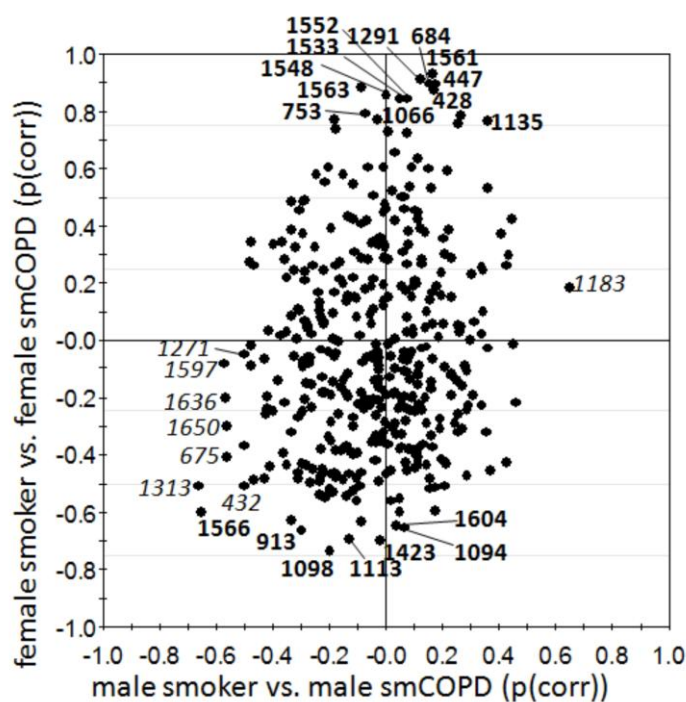
#### *Gender differences in the bronchoalveolar lavage cell proteome of patients with COPD*

In **paper II**, a proteomics analysis of the protein level alterations of alveolar macrophages in COPD was performed. The proteome of lung cells obtained by bronchoalveolar lavage (BAL) was quantified by 2D-gelelectrophoresis using fluorescent dyes (DIGE) for relative protein quantification. The study population from which the BAL samples were collected consisted of 23 non-smokers, 33 smokers with normal lung function, 15 actively smoking COPD patients and 6 ex-smoking COPD patients.

In total 404 protein "spots" with quantitative information were detected on the gels; 152 protein spots were selected for identification by mass spectrometry. Of those, 115 proteins were identified. The quantitative data was analysed by Student's t-test and multivariate statistics to identify differentially expressed proteins between the patient groups. All statistical analyses were performed comparing smoking and non-smoking individuals separately. An OPLS analysis for classification of healthy smokers and smoking COPD patients resulted in an efficient model with poor predictive power ( $Q^2=0.45$ ,  $p(\text{CV-ANOVA})=0.0003$ ). However, limiting the analysis to female subjects resulted in the identification of 19 proteins with significant protein level alterations that separated smoking subjects with normal lung function from smoking subjects diagnosed with COPD with good predictive power ( $Q^2=0.78$ ,  $p(\text{CV-ANOVA})=0.0001$ ). Corresponding analysis on the male cohort detected no proteins with significantly altered levels. A SUS-plot comparing the two models showed that these proteins were classifiers unique for the female group (Figure 5). Further, because of the gender difference observed in the literature regarding COPD phenotype, a correlation analysis of the computerised tomography (CT) data on emphysema with the proteomics results was performed. No correlation was found.

A pathway mapping of 148 protein spots with VIP >1 in the OPLS model for separation of smokers with normal lung function and smokers with COPD (females) to well-known signalling pathways showed an enrichment of 9 proteins to the lysosomal pathway ( $p < 0.0001$ ), and 7 proteins to the oxidative phosphorylation pathway ( $p < 0.0001$ ). Biologically, the findings suggest an up-regulation of the oxidative phosphorylation pathway and a down-regulation of the lysosomal pathway in early stages of COPD in this female subgroup. The alterations of the protein cathepsin B from the lysosomal pathway was confirmed by western blot analysis.

This study suggests a number of proteins and pathways involved in early stages of COPD. We conclude that search for new markers should be performed by a gender-specific approach. For future studies an increased depth in the proteome analysis could be achieved by performing a mass spectrometry based study.



**Figure 5. Analysis of shared and unique structures (SUS).** SUS-plot for comparison of the protein patterns of high importance to the OPLS models for classifying male smokers with normal lung function vs. male smokers with COPD (x-axis) and female smokers with normal lung function vs. female smokers with COPD (y-axis). This SUS-plot shows that the two models share few significant proteins. The shared structures between the compared models should have appeared at the diagonal. The variables (proteins) unique to the models appear along the x- and y-axis in the plot. Marked variables are those most influential on the models for the differentiation of male (*italic*) and female (**bold**) smokers.

### 2.3.3 Paper III

#### *Tumor proteomics by multivariate analysis on individual pathway data for characterization of vulvar cancer phenotypes*

In **paper III**, the biological aim was to increase the understanding of molecular pathways in the gynaecological cancer *vulvar squamous cell carcinoma* (VSCC) and identify the driving pathways in human papilloma virus (HPV) positive and HPV negative VSCC. A second aim was to investigate whether patient sub-groups that do or do not relapse could be discriminated. In this project we further developed a novel data analysis strategy for group level comparison of individual tumour protein expression profiles.

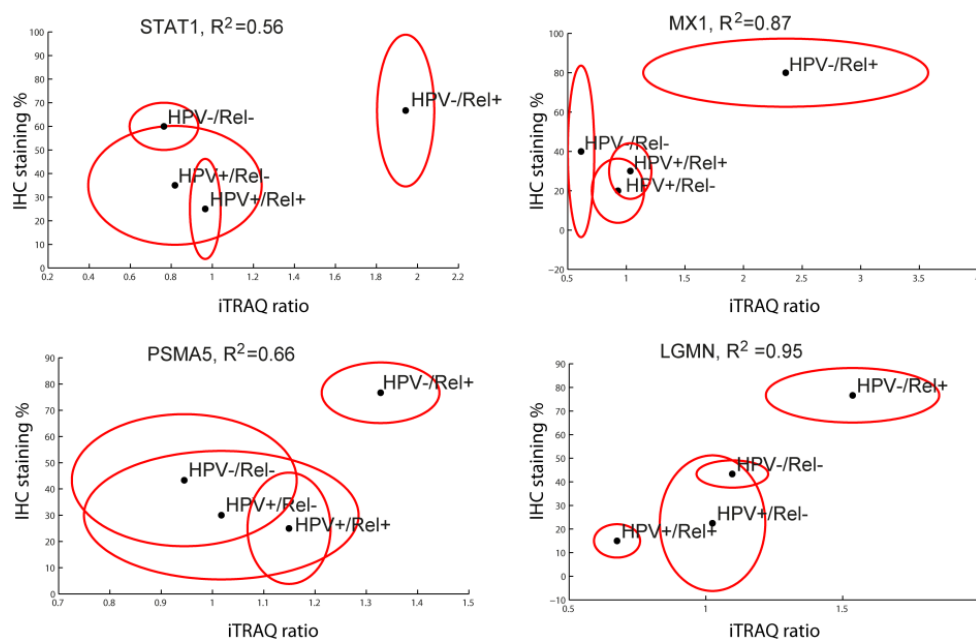
14 VSCC tumour samples (7 HPV positive and 7 HPV negative) with relapse as an independent clinical factor, were quantified by mass spectrometry using 8-plex iTRAQ labelling. In total, 1579 proteins were regarded accurately quantified and analysed further. Of the 1579 proteins, 449 were present in both iTRAQ sample sets and used for comparison of clinical sample groups defined by HPV and/or relapse status. T-test with correction for multiple testing lead to relatively few significant proteins between the groups (4 proteins for relapse and HPV classification, respectively) and high FDR values, likely because of large inter-individual variation. Due to the relatively poor overlap between the sample pools (449 proteins) we decided to also perform the individual tumour protein analysis presented in this paper.

In the individual analysis, we performed biological pathway mapping on individual tumour level. The significance measure of the matching pathway database-experimental protein data was then used in a multivariate analysis. The multivariate analysis on the pathway data was performed both unsupervised to detect clustering (PCA) and supervised (OPLS) to detect significant pathway alterations between HPV and relapse status groups. The analysis performed on overlapping proteins identified four proteins with increased expression (LGMN, MX1, STAT1 and PSMA5) associated with cancer relapse.

After validation of the mass spectrometry results by immunohistochemistry (IHC), we could single out a patient subgroup of HPV negative and relapse (HPV-/Rel+), Figure 6. These patients may be detected before treatment and might benefit from being treated as a separate group also in terms of clinical therapy. With the exception of STAT1, the proteins stained predominantly tumour cells in VSCC. This may explain why the correlation between IHC and iTRAQ data, which was strong for proteins LGMN and MX1 ( $R^2$  0.95 and 0.87, respectively), was more modest for STAT1 and PSMA5 ( $R^2$  0.56 and 0.66, respectively). Other reasons could be iTRAQ ratio underestimation (135). The strong correlation between the two methods for LGMN and MX1 however indicate that a larger clinical validation could be performed by IHC.

Two pathways were identified by both individual pathway analysis and analysis on overlapping proteins as major classifiers of relapse status: the proteasome/ubiquitin pathway and the interferon signalling pathway. These pathways included the validated proteins PSMA5, STAT1 and MX1. Both pathways are existing targets for anti-cancer therapy, although not for VSCC. Hence, the potential implementation of this data in further studies can be performed using existing drugs. Proteasome inhibitors are under evaluation for treatment against squamous cell carcinoma tumours (154-156), confirming the relevance of the proteasome in these cancers. The interferon signalling pathway with STAT1 and MX1 are known from mRNA level studies to be repressed by HPV proteins E6 and E7 (157).

This study suggests pathways and proteins significant for classifying relapse patient groups from non-relapsing patient groups. We also detected proteome level effects of the HPV virus. Further, we show that pathway fingerprinting on individual tumour level adds biological information that can strengthen a generalized protein analysis.



**Figure 6. Correlation between mass spectrometry (iTRAQ ratio) and immunohistochemistry (staining percentage) data for proteins STAT1, PSMA5, MX1, and LGMN.** The ellipses show patient subgroup classification based on IHC and MS data by sample subgroup mean (dots) and measured protein level differences within the patient groups (ellipses, +/- one standard deviation).

### 2.3.4 Paper IV

#### *Mass spectrometry based protein quantification in complex samples: the impact of labelling and precursor interference*

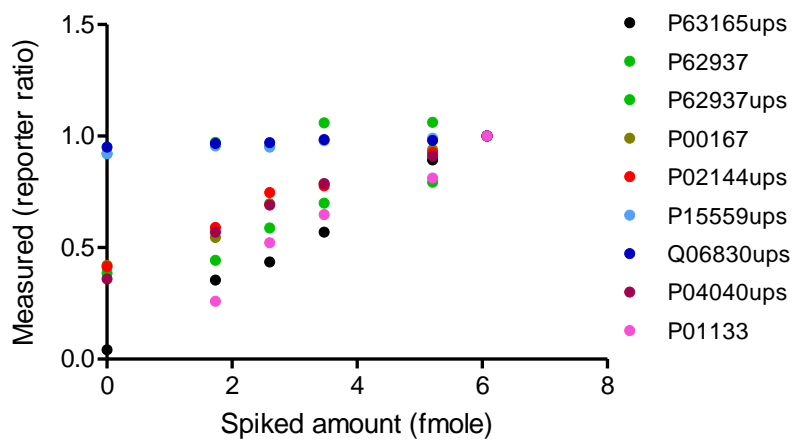
By the study in **paper IV**, the aim is to find the limits and factors affecting the quantitative accuracy and linear range for two commonly used strategies for clinical mass spectrometry-based proteomics: label-free quantification by peak area and isobaric labelling with iTRAQ and TMT. A complex biological background (mammalian cell lysate) spiked with 57 standard proteins in amounts spanning 5 orders of magnitude was used for the evaluation. To evaluate the impact of precursor mixing on the quantitative accuracy, iTRAQ and TMT labelled peptides were co-analysed. Further, pre-fractionation was performed on labelled samples by normal (pI 4-7) or narrow (pI 3.7-4.9) range isoelectric focusing, as sample complexity has an impact on precursor mixing. The narrow range IPG fractions were run using either 45 or 90 minutes LC-gradients. Label-free samples were separated on a 4 hour LC gradient. A total of 3386 proteins were identified with the label-free quantification approach, 5961 with 6-plex TMT and 4466 with 8-plex iTRAQ.

The investigation of the impact of precursor ion interference was performed by looking at the degree of contaminating reporter ions in the MS/MS fragment spectra. From that the *reporter ion interference* was calculated as contaminating reporter ion intensity/total (TMT + iTRAQ) reporter ion intensity. This was compared with the interference in the MS precursor selection window; the *isolation interference*. Our results show that precursor mixing measured by *isolation interference* and *reporter ion interference* does *not* correlate. Our explanation to these results is that this may be due to that a significant proportion of the contaminating ions in the MS precursor selection window do not give rise to reporter ions.

We further investigated the effect of precursor mixing on quantitative accuracy. From our results it appears that precursor mixing measured by *isolation interference* (MS interference) is more linked to quantitative accuracy than *reporter ion interference* (MS/MS interference). Generally, up to 30% *isolation interference* did not affect the limit of quantification (LOQ) and quantitative accuracy. But, our results indicate that precursor mixing in terms of *isolation interference* has an impact on the accuracy at lower protein levels. Another observation is that iTRAQ quantification appears less sensitive towards *isolation interference* compared to TMT.

Our results show that quantification by isobaric labels in combination with pI pre-fractionation has a lower limit of quantification (LOQ) than the label-free quantification analysis. For individual proteins quantified in a high complexity sample, the LOQ is roughly estimated to 1 femtomole for TMT, 2 femtomoles for iTRAQ, and 4 femtomoles for label-free quantification.

Based on those results, we conclude that the 6-plex TMT was more sensitive than 8-plex iTRAQ. The label-free quantification approach was least sensitive, and needs to be optimized by fractionation to reach down to the same levels as the isobaric approach as shown here.



**Figure 7. Correlation between spiked and measured amounts of protein standards added to a complex background and quantified by the isobaric label TMT.** Colouring is according to protein ID. Samples had been fractionated by narrow range isoelectric focusing, and LC gradient length 45 min.



## 2.4 GENERAL CONCLUSIONS AND FUTURE PERSPECTIVES

The overall aims of this thesis were to 1) apply quantitative proteomics for biomarker discovery in clinical samples, 2) develop data analysis workflows for the extraction of biological information from quantitative proteomics data; and 3) evaluate the potentials and limits of mass spectrometry based methods for quantitative proteomics. **The general conclusions of this thesis can thus be summarized into methodological and biological conclusions.**

### 2.4.1 Methodological conclusions

*Novel analysis workflow for quantitative proteomics data using multivariate analysis.*

The application of multivariate analysis methods on pathway data in **paper III** showed that individual protein profiles can be used for detecting subpopulations of patients without a priori knowledge of patient subgrouping. It can also be used to identify cellular signalling pathways of specific importance in pre-defined clinical groups. Advantages of analysing within the context of pathways is that it can help to relate proteome data to future targeted cancer therapies, and biomarkers can be selected from the identified pathways. Further, more proteins can be used in the analysis which is not limited to overlapping proteins. This strengthens the statistical confidence in the pathway mapping.

*Evaluation of the quantitative accuracy and limit of detection for mass spectrometry methods for clinical proteomics.* To quantify biologically relevant protein alterations, a certain depth of analysis must be reached. The current workflow in our lab includes iTRAQ-labelling and in depth-proteome profiling by narrow range isoelectric focusing followed by LC-MS/MS. With that we reach the depth to detect significant protein alterations relevant for disease as shown in **paper III**. One drawback with the label based quantification was observed in the analysis of many samples. The use of an internal reference standard allows analysis of more samples than available tags, but we observed a relatively poor overlap between LC-MS/MS runs. Clinical studies have to handle large between-sample variation within the clinical groups. This requires relatively large study groups which could be done by a label-free quantification approach. As shown in **paper IV**, we do not yet reach as deep in the proteome with the label-free setup without narrow range isoelectric focusing as with the label based with narrow range isoelectric focusing.

*Controlling bias in sample cohorts to avoid false discoveries due to confounding conditions.* Strategies to reduce bias in this study include the measurement of CO in exhaled breath (**paper I**) to verify (and control for) smoking status and measure individual differences in smoking habits. In the experimental designs of **papers II** and **III** possible confounders such as gender, age, HPV status, smoking status and disease stage were considered. For heterogeneous diseases such as COPD (**paper II**) and cancer (**paper III**), it might be necessary to select subgroups to avoid confounders and too diverse clinical groups. To limit down to studying just one subgroup, if possible, is an alternative way to find biomarkers despite heterogeneity.

### 2.4.2 Biological conclusions

Moving from experimental data on clinical samples to biological knowledge is difficult and requires a careful study design and sample selection. In addition, sensitive and accurate quantification methods are necessary to reach the low abundant proteins. Further, to be able to draw relevant conclusions from the data, the data analysis requires accurate quantitative data and robust statistical tools to detect significant protein alterations. Finally, software tools are required to interpret the protein data; both assigning proteins to biological functions by putting them in a biological context, but also by connecting proteins to each other.

*Chronic Obstructive Pulmonary disease (COPD).* In **paper II**, down-regulation of the lysosomal pathway and up-regulation of the oxidative phosphorylation pathway were linked to early stages of COPD in female smokers. These findings were unique for the female study group and indicate that future studies on COPD should consider phenotypic gender differences.

*Vulvar squamous cell carcinoma (VSCC).* The study in **paper III** contributes to the molecular understanding of VSCC and provides a number of potential proteins and pathways that could potentially result in targeted treatment of patient sub-groups. A high risk patient subgroup of HPV-negative tumours was identified based on the expression of four proteins, and the results suggest that this subgroup is characterized by an altered ubiquitin-proteasome signalling pathway.

### 2.4.3 Future perspectives

The emphasis in the clinical studies presented in this thesis lies on having in-depth proteomics coverage with many protein identities, with a trade off in the number of samples. Alternative ways to design a study is to have many patients, on the cost of analytical depth. This would increase the statistical power; which together with quantitative accuracy is key in the statistical analysis aimed for biomarker discovery. However, reducing analytical depth may lead to that the low level proteins, among which we believe disease markers are, may be missed in the analysis.

To make it practically feasible to increase the number of clinical samples, future biomarker studies requires a more high-throughput proteomics workflow. Such a proteomics workflow could be achieved by reducing the number of pI-fractions that are analysed by LC-MS/MS, which today are 72. Further, faster database searches could be achieved by using the pI to reduce the search space. This also makes it possible to extend the database to include for instance isoforms and post translational modifications.

The advances in high-throughput technologies have led to the generation of large amounts of data. The challenge is now to extract information from the data, to measure differences but still be able to draw general conclusions; and to single out biomarkers that are most relevant for further validation and in the end leads to new biological knowledge.



## ACKNOWLEDGEMENTS

I have many people to thank for contributing to this thesis; but I would like to especially thank:

Jenny Forshed, my main supervisor. It has been so nice working with you that I will have to follow your advice and highlight the essentials☺ Thank you for your enthusiasm and support; for generosity, humour and for all the things you teach me and show that I can.

Janne Lehtiö, my co-supervisor, for creating such a nice scientific environment. Thank you for welcoming me to the group, and for being an inspiring group leader with contagious ambition.

Magnus Sköld, my co supervisor. For sharing knowledge on COPD, for meetings and support, and for inviting me to the very nice Holmenkollen conferences.

I would also like to thank all co-authors and collaborators. On paper III: Gunnel Lindell, Barbro Larson, Kristina Gemzell-Danielsson, Brita Nordström-Källström and Mats Dahlberg. On papers I and II: Maxie Kohler, Åsa Wheelock, Johan Grunewald, Anders Eklund, Reza Karimi. Thank you for sharing your knowledge and for discussions around the projects.

The Cancer Proteomics Mass Spectrometry group: Lina, Helena, Lukas, Hassan, Kie, Elena, Yafeng, Henrik, Hillevi, Rui, Maria, Anna, Jorrit, Davide, Hanna and Luigi. This could not have been done without your help. Thank you for discussions and chats on various matters, for help in projects and with this thesis, for traveling to conferences together and for sharing goodies from café delta and other more exotic places. This thesis was produced in the nicest of environments!

The SciLife lab and all its research groups and administrative staff for creating royal research environment. The Lung Research Laboratory, past and present: Abraham, Bettina, Ernesto, Micke, Maria, Charlotta, Muntasir and Helena. For all the help on papers I and II: Research nurses Helene, Gunnel and Margitha; labtechnicians Benita Engwall och Benita Dahlberg. And from KBC, Susanne Becker for 2-DE expertise.

Tusen tack också till vänner och familj som bidragit med mycket på ett mer indirekt sätt☺ Bettina och Abraham, Johan L, Hanna och Emilia, Jens, Linn, Sandra och Betzy, Hanna, Johan S och Louise. Rikard. Anna. Och min familj: pappa, Linda och Jesper, Maria och Esbern, Isak och Elias.



### 3 REFERENCES

1. Harris TJ, McCormick F. The molecular pathology of cancer. *Nature reviews Clinical oncology*. 2010;7(5):251-65. Epub 2010/03/31.
2. Schwanhaussier B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al. Global quantification of mammalian gene expression control. *Nature*. 2011;473(7347):337-42. Epub 2011/05/20.
3. Maier T, Guell M, Serrano L. Correlation of mRNA and protein in complex biological samples. *FEBS Lett*. 2009;583(24):3966-73. Epub 2009/10/24.
4. Lilley KS, Razzaq A, Dupree P. Two-dimensional gel electrophoresis: recent advances in sample preparation, detection and quantitation. *Current opinion in chemical biology*. 2002;6(1):46-50. Epub 2002/02/06.
5. Nagaraj N, Kulak NA, Cox J, Neuhauser N, Mayr K, Hoerning O, 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):M111 013722. Epub 2011/10/25.
6. Branca RM, Orre L, Johansson HJ, Granholm V, Huss M, Pérez-Bercoff Å, et al. HiRIEF-LC-MS enables deep proteome coverage and unbiased proteogenomics in mouse and man. Submitted manuscript 2012.
7. Wistuba II, Gelovani JG, Jacoby JJ, Davis SE, Herbst RS. Methodological and practical challenges for personalized cancer therapies. *Nature reviews Clinical oncology*. 2011;8(3):135-41. Epub 2011/03/03.
8. Beretta L. Proteomics from the clinical perspective: many hopes and much debate. *Nat Methods*. 2007;4(10):785-6. Epub 2007/09/29.
9. Solassol J, Jacot W, Lhermitte L, Boule N, Maudelonde T, Mange A. Clinical proteomics and mass spectrometry profiling for cancer detection. *Expert review of proteomics*. 2006;3(3):311-20. Epub 2006/06/15.
10. Hanash S, Taguchi A. Application of proteomics to cancer early detection. *Cancer J*. 2011;17(6):423-8. Epub 2011/12/14.
11. Aebersold R. A stress test for mass spectrometry-based proteomics. *Nat Methods*. 2009;6(6):411-2. Epub 2009/05/19.
12. Nilsson T, Mann M, Aebersold R, Yates JR, 3rd, Bairoch A, Bergeron JJ. Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat Methods*. 2010;7(9):681-5. Epub 2010/09/02.
13. Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics*. 2005;4(10):1419-40. Epub 2005/07/13.
14. Hanash SM, Baik CS, Kallioniemi O. Emerging molecular biomarkers--blood-based strategies to detect and monitor cancer. *Nature reviews Clinical oncology*. 2011;8(3):142-50. Epub 2011/03/03.
15. Hanash S. Disease proteomics. *Nature*. 2003;422(6928):226-32. Epub 2003/03/14.
16. Liotta LA, Petricoin EF. Mass spectrometry-based protein biomarker discovery: solving the remaining challenges to reach the promise of clinical benefit. *Clinical chemistry*. 2010;56(10):1641-2. Epub 2010/07/28.
17. Poste G. Bring on the biomarkers. *Nature*. 2011;469(7329):156-7. Epub 2011/01/14.
18. Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics*. 2002;1(11):845-67. Epub 2002/12/19.
19. Landegren U, Vanelid J, Hammond M, Nong RY, Wu D, Ulleras E, et al. Opportunities for sensitive plasma proteome analysis. *Anal Chem*. 2012;84(4):1824-30. Epub 2012/01/18.
20. Cook ED, Nelson AC. Prostate cancer screening. *Current oncology reports*. 2011;13(1):57-62. Epub 2010/10/29.

21. Beck M, Schmidt A, Malmstroem J, Claassen M, Ori A, Szymborska A, et al. The quantitative proteome of a human cell line. *Mol Syst Biol*. 2011;7:549. Epub 2011/11/10.
22. Wolters DA, Washburn MP, Yates JR, 3rd. An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem*. 2001;73(23):5683-90. Epub 2002/01/05.
23. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature*. 2003;422(6928):198-207. Epub 2003/03/14.
24. Duncan MW, Aebersold R, Caprioli RM. The pros and cons of peptide-centric proteomics. *Nat Biotechnol*. 2010;28(7):659-64. Epub 2010/07/14.
25. Geiger T, Wehner A, Schaab C, Cox J, Mann M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol Cell Proteomics*. 2012;11(3):M111 014050. Epub 2012/01/27.
26. Domon B, Aebersold R. Options and considerations when selecting a quantitative proteomics strategy. *Nat Biotechnol*. 2010;28(7):710-21. Epub 2010/07/14.
27. Steen H, Mann M. The ABC's (and XYZ's) of peptide sequencing. *Nature reviews Molecular cell biology*. 2004;5(9):699-711. Epub 2004/09/02.
28. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. 1999;20(18):3551-67. Epub 1999/12/28.
29. Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods*. 2007;4(10):787-97. Epub 2007/09/29.
30. Nesvizhskii AI. Protein identification by tandem mass spectrometry and sequence database searching. *Methods Mol Biol*. 2007;367:87-119. Epub 2006/12/23.
31. Rappsilber J, Mann M. What does it mean to identify a protein in proteomics? *Trends in biochemical sciences*. 2002;27(2):74-8. Epub 2002/02/20.
32. Forshed J, Johansson HJ, Pernemalm M, Branca RM, Sandberg A, Lehtio J. Enhanced information output from shotgun proteomics data by protein quantification and peptide quality control (PQPQ). *Mol Cell Proteomics*. 2011;10(10):M111 010264. Epub 2011/07/08.
33. Ong SE, Mann M. Mass spectrometry-based proteomics turns quantitative. *Nature chemical biology*. 2005;1(5):252-62. Epub 2006/01/13.
34. Bachi A, Bonaldi T. Quantitative proteomics as a new piece of the systems biology puzzle. *Journal of proteomics*. 2008;71(3):357-67. Epub 2008/07/22.
35. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and bioanalytical chemistry*. 2007;389(4):1017-31. Epub 2007/08/02.
36. Liu H, Sadygov RG, Yates JR, 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem*. 2004;76(14):4193-201. Epub 2004/07/16.
37. Chelius D, Bondarenko PV. Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *J Proteome Res*. 2002;1(4):317-23. Epub 2003/03/21.
38. Wiener MC, Sachs JR, Deyanova EG, Yates NA. Differential mass spectrometry: a label-free LC-MS method for finding significant differences in complex peptide and protein mixtures. *Anal Chem*. 2004;76(20):6085-96. Epub 2004/10/16.
39. Luber CA, Cox J, Lauterbach H, Fancke B, Selbach M, Tschopp J, et al. Quantitative proteomics reveals subset-specific viral recognition in dendritic cells. *Immunity*. 2010;32(2):279-89. Epub 2010/02/23.
40. Griffin NM, Yu J, Long F, Oh P, Shore S, Li Y, et al. Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat Biotechnol*. 2010;28(1):83-9. Epub 2009/12/17.
41. Mueller LN, Rinner O, Schmidt A, Letarte S, Bodenmiller B, Brusniak MY, et al. SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics*. 2007;7(19):3470-80. Epub 2007/08/30.
42. Leitner A, Lindner W. Chemistry meets proteomics: the use of chemical tagging reactions for MS-based proteomics. *Proteomics*. 2006;6(20):5418-34. Epub 2006/09/15.

43. Gouw JW, Krijgsveld J, Heck AJ. Quantitative proteomics by metabolic labeling of model organisms. *Mol Cell Proteomics*. 2010;9(1):11-24. Epub 2009/12/04.
44. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, 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):376-86. Epub 2002/07/16.
45. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*. 2008;26(12):1367-72. Epub 2008/11/26.
46. Ow SY, Cardona T, Taton A, Magnuson A, Lindblad P, Stensjo K, et al. Quantitative shotgun proteomics of enriched heterocysts from *Nostoc* sp. PCC 7120 using 8-plex isobaric peptide tags. *J Proteome Res*. 2008;7(4):1615-28. Epub 2008/02/23.
47. Engmann O, Campbell J, Ward M, Giese KP, Thompson AJ. Comparison of a protein-level and peptide-level labeling strategy for quantitative proteomics of synaptosomes using isobaric tags. *J Proteome Res*. 2010;9(5):2725-33. Epub 2010/03/12.
48. Dayon L, Hainard A, Licker V, Turck N, Kuhn K, Hochstrasser DF, et al. Relative quantification of proteins in human cerebrospinal fluids by MS/MS using 6-plex isobaric tags. *Anal Chem*. 2008;80(8):2921-31. Epub 2008/03/04.
49. Dayon L, Turck N, Kienle S, Schulz-Knappe P, Hochstrasser DF, Scherl A, et al. Isobaric tagging-based selection and quantitation of cerebrospinal fluid tryptic peptides with reporter calibration curves. *Anal Chem*. 2010;82(3):848-58. Epub 2010/01/12.
50. Werner T, Becher I, Sweetman G, Doce C, Savitski MM, Bantscheff M. High-Resolution Enabled TMT 8-plexing. *Anal Chem*. 2012;84(16):7188-94. Epub 2012/08/14.
51. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*. 2000;25(1):25-9. Epub 2000/05/10.
52. Thomas S, Bonchev D. A survey of current software for network analysis in molecular biology. *Human genomics*. 2010;4(5):353-60. Epub 2010/07/24.
53. Malik R, Dulla K, Nigg EA, Korner R. From proteome lists to biological impact--tools and strategies for the analysis of large MS data sets. *Proteomics*. 2010;10(6):1270-83. Epub 2010/01/16.
54. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell*. 2000;100(1):57-70. Epub 2000/01/27.
55. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144(5):646-74. Epub 2011/03/08.
56. Levine AJ. p53, the cellular gatekeeper for growth and division. *Cell*. 1997;88(3):323-31. Epub 1997/02/07.
57. Yap TA, Gerlinger M, Futreal PA, Pusztai L, Swanton C. Intratumor heterogeneity: seeing the wood for the trees. *Science translational medicine*. 2012;4(127):127ps10. Epub 2012/03/31.
58. Kulasingam V, Diamandis EP. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. *Nature clinical practice Oncology*. 2008;5(10):588-99. Epub 2008/08/13.
59. Sawyers CL. The cancer biomarker problem. *Nature*. 2008;452(7187):548-52. Epub 2008/04/04.
60. Dancey JE, Chen HX. Strategies for optimizing combinations of molecularly targeted anticancer agents. *Nature reviews Drug discovery*. 2006;5(8):649-59. Epub 2006/08/03.
61. zur Hausen H. Papillomaviruses in the causation of human cancers - a brief historical account. *Virology*. 2009;384(2):260-5. Epub 2009/01/13.
62. Moody CA, Laimins LA. Human papillomavirus oncoproteins: pathways to transformation. *Nat Rev Cancer*. 2010;10(8):550-60. Epub 2010/07/02.
63. Tindle RW. Immune evasion in human papillomavirus-associated cervical cancer. *Nat Rev Cancer*. 2002;2(1):59-65. Epub 2002/03/21.
64. McCluggage WG. Recent developments in vulvovaginal pathology. *Histopathology*. 2009;54(2):156-73. Epub 2008/07/19.

65. Brandenberger AW, Rudlinger R, Hanggi W, Bersinger NA, Dreher E. Detection of human papillomavirus in vulvar carcinoma. A study by in situ hybridisation. *Arch Gynecol Obstet.* 1992;252(1):31-5. Epub 1992/01/01.
66. Iwasawa A, Nieminen P, Lehtinen M, Paavonen J. Human papillomavirus in squamous cell carcinoma of the vulva by polymerase chain reaction. *Obstet Gynecol.* 1997;89(1):81-4. Epub 1997/01/01.
67. Lerma E, Matias-Guiu X, Lee SJ, Prat J. Squamous cell carcinoma of the vulva: study of ploidy, HPV, p53, and pRb. *Int J Gynecol Pathol.* 1999;18(3):191-7. Epub 2002/07/02.
68. Pinto AP, Schlecht NF, Pintos J, Kaiano J, Franco EL, Crum CP, et al. Prognostic significance of lymph node variables and human papillomavirus DNA in invasive vulvar carcinoma. *Gynecol Oncol.* 2004;92(3):856-65. Epub 2004/02/27.
69. Skapa P, Zamecnik J, Hamsikova E, Salakova M, Smahelova J, Jandova K, et al. Human papillomavirus (HPV) profiles of vulvar lesions: possible implications for the classification of vulvar squamous cell carcinoma precursors and for the efficacy of prophylactic HPV vaccination. *Am J Surg Pathol.* 2007;31(12):1834-43. Epub 2007/11/29.
70. Knopp S, Nesland JM, Trope C, Holm R. p14ARF, a prognostic predictor in HPV-negative vulvar carcinoma. *Am J Clin Pathol.* 2006;126(2):266-76. Epub 2006/08/08.
71. Madsen BS, Jensen HL, van den Brule AJ, Wohlfahrt J, Frisch M. Risk factors for invasive squamous cell carcinoma of the vulva and vagina--population-based case-control study in Denmark. *Int J Cancer.* 2008;122(12):2827-34. Epub 2008/03/19.
72. Sutton BC, Allen RA, Moore WE, Dunn ST. Distribution of human papillomavirus genotypes in invasive squamous carcinoma of the vulva. *Mod Pathol.* 2008;21(3):345-54. Epub 2008/01/15.
73. Lindell G, Nasman A, Jonsson C, Ehrsson RJ, Jacobsson H, Danielsson KG, et al. Presence of human papillomavirus (HPV) in vulvar squamous cell carcinoma (VSCC) and sentinel node. *Gynecol Oncol.* 2010;117(2):312-6. Epub 2010/02/09.
74. Monk BJ, Burger RA, Lin F, Parham G, Vasilev SA, Wilczynski SP. Prognostic significance of human papillomavirus DNA in vulvar carcinoma. *Obstet Gynecol.* 1995;85(5 Pt 1):709-15. Epub 1995/05/01.
75. van de Nieuwenhof HP, van Kempen LC, de Hullu JA, Bekkers RL, Bulten J, Melchers WJ, et al. The etiologic role of HPV in vulvar squamous cell carcinoma fine tuned. *Cancer Epidemiol Biomarkers Prev.* 2009;18(7):2061-7. Epub 2009/07/02.
76. Alonso I, Fuste V, del Pino M, Castillo P, Torne A, Fuste P, et al. Does human papillomavirus infection imply a different prognosis in vulvar squamous cell carcinoma? *Gynecol Oncol.* 2011;122(3):509-14. Epub 2011/06/10.
77. Ansink AC, Sie-Go DM, van der Velden J, Sijmons EA, de Barros Lopes A, Monaghan JM, et al. Identification of sentinel lymph nodes in vulvar carcinoma patients with the aid of a patent blue V injection: a multicenter study. *Cancer.* 1999;86(4):652-6. Epub 1999/08/10.
78. Hording U, Daugaard S, Junge J, Lundvall F. Human papillomaviruses and multifocal genital neoplasia. *Int J Gynecol Pathol.* 1996;15(3):230-4. Epub 1996/07/01.
79. Gray HJ. Advances in vulvar and vaginal cancer treatment. *Gynecol Oncol.* 2010;118(1):3-5. Epub 2010/05/18.
80. Horowitz NS, Olawaiye AB, Borger DR, Growdon WB, Krasner CN, Matulonis UA, et al. Phase II trial of erlotinib in women with squamous cell carcinoma of the vulva. *Gynecol Oncol.* 2012;127(1):141-6. Epub 2012/07/04.
81. Sopori M. Effects of cigarette smoke on the immune system. *Nature reviews Immunology.* 2002;2(5):372-7. Epub 2002/05/30.
82. van der Vaart H, Postma DS, Timens W, ten Hacken NH. Acute effects of cigarette smoke on inflammation and oxidative stress: a review. *Thorax.* 2004;59(8):713-21. Epub 2004/07/30.
83. Wehlin L, Lofdahl M, Lundahl J, Skold M. Reduced intracellular oxygen radical production in whole blood leukocytes from COPD patients and asymptomatic smokers. *Chest.* 2005;128(4):2051-8. Epub 2005/10/21.



84. Kohansal R, Martinez-Camblor P, Agusti A, Buist AS, Mannino DM, Soriano JB. The natural history of chronic airflow obstruction revisited: an analysis of the Framingham offspring cohort. *Am J Respir Crit Care Med*. 2009;180(1):3-10. Epub 2009/04/04.
85. Barnes PJ, Shapiro SD, Pauwels RA. Chronic obstructive pulmonary disease: molecular and cellular mechanisms. *The European respiratory journal : official journal of the European Society for Clinical Respiratory Physiology*. 2003;22(4):672-88. Epub 2003/10/30.
86. Hogg JC. Pathophysiology of airflow limitation in chronic obstructive pulmonary disease. *Lancet*. 2004;364(9435):709-21. Epub 2004/08/25.
87. From the Global Strategy for the Diagnosis, Management and Prevention of COPD, Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2011. Available from: <http://www.goldcopd.org/>. (accessed Sept 19, 2012).
88. Hogg JC, Chu F, Utokaparch S, Woods R, Elliott WM, Buzatu L, et al. The nature of small-airway obstruction in chronic obstructive pulmonary disease. *N Engl J Med*. 2004;350(26):2645-53. Epub 2004/06/25.
89. Russell RE, Thorley A, Culpitt SV, Dodd S, Donnelly LE, Demattos C, et al. Alveolar macrophage-mediated elastolysis: roles of matrix metalloproteinases, cysteine, and serine proteases. *American journal of physiology Lung cellular and molecular physiology*. 2002;283(4):L867-73. Epub 2002/09/13.
90. Culpitt SV, Rogers DF, Shah P, De Matos C, Russell RE, Donnelly LE, et al. Impaired inhibition by dexamethasone of cytokine release by alveolar macrophages from patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 2003;167(1):24-31. Epub 2002/10/31.
91. Barnes PJ, Ito K, Adcock IM. Corticosteroid resistance in chronic obstructive pulmonary disease: inactivation of histone deacetylase. *Lancet*. 2004;363(9410):731-3. Epub 2004/03/06.
92. Lundback B, Lindberg A, Lindstrom M, Ronmark E, Jonsson AC, Jonsson E, et al. Not 15 but 50% of smokers develop COPD?--Report from the Obstructive Lung Disease in Northern Sweden Studies. *Respiratory medicine*. 2003;97(2):115-22. Epub 2003/02/18.
93. McCloskey SC, Patel BD, Hinchliffe SJ, Reid ED, Wareham NJ, Lomas DA. Siblings of patients with severe chronic obstructive pulmonary disease have a significant risk of airflow obstruction. *Am J Respir Crit Care Med*. 2001;164(8 Pt 1):1419-24. Epub 2001/11/13.
94. Hunninghake GM, Cho MH, Tesfaigzi Y, Soto-Quiros ME, Avila L, Lasky-Su J, et al. MMP12, lung function, and COPD in high-risk populations. *N Engl J Med*. 2009;361(27):2599-608. Epub 2009/12/19.
95. Abbey DE, Burchette RJ, Knutsen SF, McDonnell WF, Lebowitz MD, Enright PL. Long-term particulate and other air pollutants and lung function in nonsmokers. *Am J Respir Crit Care Med*. 1998;158(1):289-98. Epub 1998/07/09.
96. Menezes AM, Hallal PC, Perez-Padilla R, Jardim JR, Muino A, Lopez MV, et al. Tuberculosis and airflow obstruction: evidence from the PLATINO study in Latin America. *The European respiratory journal : official journal of the European Society for Clinical Respiratory Physiology*. 2007;30(6):1180-5. Epub 2007/09/07.
97. Eisner MD, Balmes J, Katz PP, Trupin L, Yelin EH, Blanc PD. Lifetime environmental tobacco smoke exposure and the risk of chronic obstructive pulmonary disease. *Environmental health : a global access science source*. 2005;4(1):7. Epub 2005/05/14.
98. Brostrom EB, Thunqvist P, Adenfelt G, Borling E, Katz-Salamon M. Obstructive lung disease in children with mild to severe BPD. *Respiratory medicine*. 2010;104(3):362-70. Epub 2009/11/13.
99. Mannino DM, Homa DM, Akinbami LJ, Ford ES, Redd SC. Chronic obstructive pulmonary disease surveillance--United States, 1971-2000. *MMWR Surveill Summ*. 2002;51(6):1-16. Epub 2002/08/30.
100. Foreman MG, Zhang L, Murphy J, Hansel NN, Make B, Hokanson JE, et al. Early-onset chronic obstructive pulmonary disease is associated with female sex, maternal factors, and African American race in the COPD Gene Study. *Am J Respir Crit Care Med*. 2011;184(4):414-20. Epub 2011/05/13.

101. Lopez Varela MV, Montes de Oca M, Halbert RJ, Muino A, Perez-Padilla R, Talamo C, et al. Sex-related differences in COPD in five Latin American cities: the PLATINO study. *The European respiratory journal : official journal of the European Society for Clinical Respiratory Physiology*. 2010;36(5):1034-41. Epub 2010/04/10.
102. Sorheim IC, Johannessen A, Gulsvik A, Bakke PS, Silverman EK, DeMeo DL. Gender differences in COPD: are women more susceptible to smoking effects than men? *Thorax*. 2010;65(6):480-5. Epub 2010/06/05.
103. Martinez FJ, Curtis JL, Sciurba F, Mumford J, Giardino ND, Weinmann G, et al. Sex differences in severe pulmonary emphysema. *Am J Respir Crit Care Med*. 2007;176(3):243-52. Epub 2007/04/14.
104. WHO. World health statistics 2008.  
[http://www.who.int/gho/publications/world\\_health\\_statistics/EN\\_WHS08\\_Full.pdf](http://www.who.int/gho/publications/world_health_statistics/EN_WHS08_Full.pdf)  
(accessed Sept 19, 2012).
105. Pistolesi M, Camiciottoli G, Paoletti M, Marmai C, Lavorini F, Meoni E, et al. Identification of a predominant COPD phenotype in clinical practice. *Respiratory medicine*. 2008;102(3):367-76. Epub 2008/02/06.
106. Decramer M, Rennard S, Troosters T, Mapel DW, Giardino N, Mannino D, et al. COPD as a lung disease with systemic consequences--clinical impact, mechanisms, and potential for early intervention. *Copd*. 2008;5(4):235-56. Epub 2008/08/02.
107. Celli BR, Cote CG, Marin JM, Casanova C, Montes de Oca M, Mendez RA, et al. The body-mass index, airflow obstruction, dyspnea, and exercise capacity index in chronic obstructive pulmonary disease. *N Engl J Med*. 2004;350(10):1005-12. Epub 2004/03/05.
108. Calverley PM, Anderson JA, Celli B, Ferguson GT, Jenkins C, Jones PW, et al. Salmeterol and fluticasone propionate and survival in chronic obstructive pulmonary disease. *N Engl J Med*. 2007;356(8):775-89. Epub 2007/02/23.
109. Pauwels RA, Buist AS, Calverley PM, Jenkins CR, Hurd SS. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. NHLBI/WHO Global Initiative for Chronic Obstructive Lung Disease (GOLD) Workshop summary. *Am J Respir Crit Care Med*. 2001;163(5):1256-76. Epub 2001/04/24.
110. Olsen HH, Grunewald J, Tornling G, Skold CM, Eklund A. Bronchoalveolar lavage results are independent of season, age, gender and collection site. *PLoS One*. 2012;7(8):e43644. Epub 2012/09/07.
111. Karimi R, Tornling G, Grunewald J, Eklund A, Skold CM. Cell recovery in bronchoalveolar lavage fluid in smokers is dependent on cumulative smoking history. *PLoS One*. 2012;7(3):e34232. Epub 2012/04/06.
112. Eriksson H, Lengqvist J, Hedlund J, Uhlen K, Orre LM, Bjellqvist B, et al. Quantitative membrane proteomics applying narrow range peptide isoelectric focusing for studies of small cell lung cancer resistance mechanisms. *Proteomics*. 2008;8(15):3008-18.
113. Cargile BJ, Stephenson JL, Jr. An alternative to tandem mass spectrometry: isoelectric point and accurate mass for the identification of peptides. *Anal Chem*. 2004;76(2):267-75. Epub 2004/01/15.
114. Cargile BJ, Bundy JL, Freeman TW, Stephenson JL, Jr. Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification. *J Proteome Res*. 2004;3(1):112-9. Epub 2004/03/05.
115. Lengqvist J, Uhlen K, Lehtio J. iTRAQ compatibility of peptide immobilized pH gradient isoelectric focusing. *Proteomics*. 2007;7(11):1746-52.
116. Sevinsky JR, Cargile BJ, Bunger MK, Meng F, Yates NA, Hendrickson RC, et al. Whole genome searching with shotgun proteomic data: applications for genome annotation. *J Proteome Res*. 2008;7(1):80-8. Epub 2007/12/08.
117. Gorg A, Obermaier C, Boguth G, Harder A, Scheibe B, Wildgruber R, et al. The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*. 2000;21(6):1037-53. Epub 2000/04/29.
118. Unlu M, Morgan ME, Minden JS. Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis*. 1997;18(11):2071-7. Epub 1998/01/07.

119. Alban A, David SO, Bjorkesten L, Andersson C, Sloge E, Lewis S, et al. A novel experimental design for comparative two-dimensional gel analysis: two-dimensional difference gel electrophoresis incorporating a pooled internal standard. *Proteomics*. 2003;3:36-44.
120. Shaw J, Rowlinson R, Nickson J, Stone T, Sweet A, Williams K, et al. Evaluation of saturation labelling two-dimensional difference gel electrophoresis fluorescent dyes. *Proteomics* 2003Jul;3(7):1181-95. 2003;3:1181-95.
121. Gygi SP, Corthals GL, Zhang Y, Rochon Y, Aebersold R. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc Natl Acad Sci U S A*. 2000;97(17):9390-5. Epub 2000/08/02.
122. Han MJ, Herlyn M, Fisher AB, Speicher DW. Microscale solution IEF combined with 2-D DIGE substantially enhances analysis depth of complex proteomes such as mammalian cell and tissue extracts. *Electrophoresis*. 2008;29(3):695-705. Epub 2008/01/12.
123. Sitek B, Sipos B, Pfeiffer K, Grzendowski M, Poschmann G, Hawranke E, et al. Establishment of "one-piece" large-gel 2-DE for high-resolution analysis of small amounts of sample using difference gel electrophoresis saturation labelling. *Analytical and bioanalytical chemistry*. 2008;391(1):361-5. Epub 2008/04/02.
124. Perry RH, Cooks RG, Noll RJ. Orbitrap mass spectrometry: instrumentation, ion motion and applications. *Mass spectrometry reviews*. 2008;27(6):661-99. Epub 2008/08/08.
125. Makarov A, Denisov E, Kholomeev A, Balschun W, Lange O, Strupat K, et al. Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal Chem*. 2006;78(7):2113-20. Epub 2006/04/04.
126. Dayon L, Pasquarello C, Hoogland C, Sanchez JC, Scherl A. Combining low- and high-energy tandem mass spectra for optimized peptide quantification with isobaric tags. *Journal of proteomics*. 2009.
127. Boja ES, Phillips D, French SA, Harris RA, Balaban RS. Quantitative mitochondrial phosphoproteomics using iTRAQ on an LTQ-Orbitrap with high energy collision dissociation. *J Proteome Res*. 2009;8(10):4665-75. Epub 2009/08/22.
128. Kocher T, Pichler P, Schutzbier M, Stingl C, Kaul A, Teucher N, et al. High precision quantitative proteomics using iTRAQ on an LTQ Orbitrap: a new mass spectrometric method combining the benefits of all. *J Proteome Res*. 2009;8(10):4743-52. Epub 2009/08/12.
129. Phanstiel D, Unwin R, McAlister GC, Coon JJ. Peptide quantification using 8-plex isobaric tags and electron transfer dissociation tandem mass spectrometry. *Anal Chem*. 2009;81(4):1693-8. Epub 2009/01/22.
130. Phanstiel D, Zhang Y, Marto JA, Coon JJ. Peptide and protein quantification using iTRAQ with electron transfer dissociation. *J Am Soc Mass Spectrom*. 2008;19(9):1255-62. Epub 2008/07/16.
131. Bantscheff M, Boesche M, Eberhard D, Matthieson T, Sweetman G, Kuster B. Robust and sensitive iTRAQ quantification on an LTQ Orbitrap mass spectrometer. *Mol Cell Proteomics*. 2008;7(9):1702-13. Epub 2008/05/31.
132. Savitski MM, Fischer F, Mathieson T, Sweetman G, Lang M, Bantscheff M. Targeted data acquisition for improved reproducibility and robustness of proteomic mass spectrometry assays. *J Am Soc Mass Spectrom*. 2010;21(10):1668-79. Epub 2010/02/23.
133. Michalski A, Cox J, Mann M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J Proteome Res*. 2011;10(4):1785-93. Epub 2011/02/12.
134. Houel S, Abernathy R, Renganathan K, Meyer-Arendt K, Ahn NG, Old WM. Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *J Proteome Res*. 2010;9(8):4152-60. Epub 2010/06/29.
135. Karp NA, Huber W, Sadowski PG, Charles PD, Hester SV, Lilley KS. Addressing accuracy and precision issues in iTRAQ quantitation. *Mol Cell Proteomics*. 2010;9(9):1885-97. Epub 2010/04/13.
136. Ow SY, Salim M, Noirel J, Evans C, Rehman I, Wright PC. iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly". *J Proteome Res*. 2009;8(11):5347-55. Epub 2009/09/17.

137. Christoforou A, Lilley KS. Taming the isobaric tagging elephant in the room in quantitative proteomics. *Nat Methods*. 2011;8(11):911-3. Epub 2011/11/01.
138. Ow SY, Salim M, Noirel J, Evans C, Wright PC. Minimising iTRAQ ratio compression through understanding LC-MS elution dependence and high-resolution HILIC fractionation. *Proteomics*. 2011;11(11):2341-6. Epub 2011/05/07.
139. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, et al. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics*. 2004;3(12):1154-69.
140. Old WM, Meyer-Arendt K, Aveline-Wolf L, Pierce KG, Mendoza A, Sevinsky JR, et al. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol Cell Proteomics*. 2005;4(10):1487-502. Epub 2005/06/28.
141. Wang W, Zhou H, Lin H, Roy S, Shaler TA, Hill LR, et al. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal Chem*. 2003;75(18):4818-26. Epub 2003/12/17.
142. Kulthra K, Nilsson A, Scholz B, Rossbach UL, Falth M, Andren PE. Development and evaluation of normalization methods for label-free relative quantification of endogenous peptides. *Mol Cell Proteomics*. 2009;8(10):2285-95. Epub 2009/07/15.
143. Callister SJ, Barry RC, Adkins JN, Johnson ET, Qian WJ, Webb-Robertson BJ, et al. Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J Proteome Res*. 2006;5(2):277-86. Epub 2006/02/07.
144. Lu P, Vogel C, Wang R, Yao X, Marcotte EM. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol*. 2007;25(1):117-24. Epub 2006/12/26.
145. Reiter L, Claassen M, Schrimpf SP, Jovanovic M, Schmidt A, Buhmann JM, et al. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol Cell Proteomics*. 2009;8(11):2405-17. Epub 2009/07/18.
146. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*. 2001;98(9):5116-21. Epub 2001/04/20.
147. Roxas BAP, Li QB. Significance analysis of microarray for relative quantitation of LC/MS data in proteomics. *Bmc Bioinformatics*. 2008;9.
148. Uhlen M, Bjorling E, Agaton C, Szigartyo CA, Amini B, Andersen E, et al. A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol Cell Proteomics*. 2005;4(12):1920-32. Epub 2005/08/30.
149. Sato S, Nishimura K, Koyama H, Tsukino M, Oga T, Hajiuro T, et al. Optimal cutoff level of breath carbon monoxide for assessing smoking status in patients with asthma and COPD. *Chest*. 2003;124(5):1749-54. Epub 2003/11/08.
150. Andersson MF, Moller AM. Assessment of carbon monoxide values in smokers: a comparison of carbon monoxide in expired air and carboxyhaemoglobin in arterial blood. *European journal of anaesthesiology*. 2010;27(9):812-8. Epub 2010/06/04.
151. Kharitonov SA, Barnes PJ. Exhaled markers of pulmonary disease. *Am J Respir Crit Care Med*. 2001;163(7):1693-722. Epub 2001/06/13.
152. Deller A, Stenz R, Forstner K, Konrad F. [The elimination of carboxyhemoglobin--gender-specific and circadian effects]. *Infusionstherapie und Transfusionsmedizin*. 1992;19(3):121-6. Epub 1992/06/01. Die Elimination von Kohlenmonoxydhamoglobin--Geschlechtsspezifische und zirkadiane Einflüsse.
153. Koczulla AR, Noeske S, Herr C, Jorres RA, Rommelt H, Vogelmeier C, et al. Acute and chronic effects of smoking on inflammation markers in exhaled breath condensate in current smokers. *Respiration; international review of thoracic diseases*. 2010;79(1):61-7. Epub 2009/10/10.
154. Latonen L, Moore HM, Bai B, Jaamaa S, Laiho M. Proteasome inhibitors induce nucleolar aggregation of proteasome target proteins and polyadenylated RNA by altering ubiquitin availability. *Oncogene*. 2010. Epub 2010/10/20.
155. Navon A, Ciechanover A. The 26 S proteasome: from basic mechanisms to drug targeting. *J Biol Chem*. 2009;284(49):33713-8. Epub 2009/10/09.

156. Strome SE, Kawakami K, Alejandro D, Voss S, Kasperbauer JL, Salomao D, et al. Interleukin 4 receptor-directed cytotoxin therapy for human head and neck squamous cell carcinoma in animal models. *Clin Cancer Res.* 2002;8(1):281-6. Epub 2002/01/22.
157. Chang YE, Laimins LA. Microarray analysis identifies interferon-inducible genes and Stat-1 as major transcriptional targets of human papillomavirus type 31. *J Virol.* 2000;74(9):4174-82. Epub 2001/02/07.