Turun yliopisto
University of Turku

TISSUE PROTEOMES:

Quantitative Mass Spectrometry of Murine Liver and Ovarian Endometrioma

Anni Vehmas

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
Sarja - ser. D osa - tom. 1236 | Medica - Odontologica | Turku 2016

University of Turku

Faculty of Medicine
Department of Physiology
Turku Doctoral Programme of Molecular Medicine, TuDMM
National Doctoral Program in Informational and Structural Biology, ISB
Turku Centre for Biotechnology

Supervised by

Professor Matti Poutanen, PhD
Department of Physiology
Institute of Biomedicine
University of Turku
Turku, Finland

Professor Garry Corthals, PhD
Van 't Hoff Institute for Molecular Sciences
University of Amsterdam
Amsterdam, the Netherlands

Reviewed by

Docent Hannu Koistinen, DSc
Department of Clinical Chemistry
Medicum
University of Helsinki
Helsinki, Finland

Charles Pineau, PhD
Institute de recherche en santé,
environnement et travail
University of Rennes 1
the French National Institute of Health and
Medical Research
Rennes, France

Opponent

Professor György Marko-Varga, PhD
Department of Biomedical Engineering
Biomedical Center
Lund University
Lund, Sweden

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

ISBN 978-951-29-6524-3 (PRINT)
ISBN 978-951-29-6525-0 (PDF)
ISSN 0355-9483 (Print)
ISSN 2343-3213 (Online)
Painosalama Oy - Turku, Finland 2016

ABSTRACT

Anni Vehmas

Tissue Proteomes: Quantitative Mass Spectrometry of Murine Liver and Ovarian Endometrioma

Department of Physiology, Institute of Biomedicine and Turku Centre for Biotechnology, University of Turku, Turku, Finland
Annales Universitatis Turkuensis
Turku 2016

A human genome contains more than 20 000 protein-encoding genes. A human proteome, instead, has been estimated to be much more complex and dynamic. The most powerful tool to study proteins today is mass spectrometry (MS). MS based proteomics is based on the measurement of the masses of charged peptide ions in a gas-phase. The peptide amino acid sequence can be deduced, and matching proteins can be found, using software to correlate MS-data with sequence database information. Quantitative proteomics allow the estimation of the absolute or relative abundance of a certain protein in a sample. The label-free quantification methods use the intrinsic MS-peptide signals in the calculation of the quantitative values enabling the comparison of peptide signals from numerous patient samples.

In this work, a quantitative MS methodology was established to study aromatase overexpressing (AROM+) male mouse liver and ovarian endometriosis tissue samples. The workflow of label-free quantitative proteomics was optimized in terms of sensitivity and robustness, allowing the quantification of 1500 proteins with a low coefficient of variance in both sample types. Additionally, five statistical methods were evaluated for the use with label-free quantitative proteomics data.

The proteome data was integrated with other omics datasets, such as mRNA microarray and metabolite data sets. As a result, an altered lipid metabolism in liver was discovered in male AROM+ mice. The results suggest a reduced beta oxidation of long chain phospholipids in the liver and increased levels of pro-inflammatory fatty acids in the circulation in these mice. Conversely, in the endometriosis tissues, a set of proteins highly specific for ovarian endometrioma were discovered, many of which were under the regulation of the growth factor TGF-β1. This finding supports subsequent biomarker verification in a larger number of endometriosis patient samples.

Keywords: Tissue, mass spectrometry, proteomics, biomarker, metabolomics, transcriptomics, liver, aromatase P450, mouse, endometriosis, ovarian endometrioma, statistics, Reproducibility-optimized test statistic (ROTS)

TIIVISTELMÄ

Anni Vehmas

Kudosten proteomit: Hiiren maksa ja munasarjan endometrioosi kvantitatiivisen massaspektrometrian menetelmin

Fysiologia, Biolääketieteen laitos ja Turun biotekniikan keskus, Turun yliopisto, Turku
Annales Universitatis Turkuensis
Turku 2016

Ihmisen genomin on arvioitu sisältävän yli 20 000 proteiinia koodaavaa geeniä, mutta proteiinin kokonaismäärän on kuitenkin arvioitu olevan paljon suurempi. Tehokkain menetelmä kaikkien kudoksessa ilmentyvien proteiinien (proteomin) tutkimiseksi on massaspektrometria (MS). MS-proteomiikka perustuu peptidi-ionien massojen mittaamiseen kaasufaasissa. Näiden mittausten perusteella peptidin aminohappojärjestys voidaan määrittää ja peptidiä vastaava proteiini voidaan tunnistaa genomitietokannoista. Kvantitatiivisen proteomiikan menetelmät mahdollistavat tietyn proteiinin absoluuttisen tai suhteellinen määrän osoittamisen näytteestä.

Kvantitatiivista proteomiikkaa sovellettiin tässä työssä aromataasientsyymiä ylikspressoivien (AROM +) uroshiirien maksan, ja endometrioositautia sairastavien naisten munasarjan endometrioosikudoksen tutkimuksessa. Menetelmä optimoitiin parhaan herkkyyden ja toistettavuuden saavuttamiseksi, ja lopulta noin 1500 proteiinia pystyttiin kvantifioimaan hyvällä toistettavuudella molemmissa näytetyypeissä. Työssä testattiin myös viiden eri tilastollisen menetelmän soveltuvuutta leimavapaalle kvantitatiiviselle proteomiikka-aineistolle.

Proteomiikan menetelmien lisäksi AROM+ uroshiirten maksan toimintaa tutkittiin käyttäen transkriptomiikan ja metabolomiikan menetelmiä. Tulokset osoittivat, että AROM+ uroshiirten maksan rasva-aineenvaihdunta oli muuttunut. Erityisesti pitkäketjuisten fosfolipidien β-oksidaatio oli vähentynyt maksassa ja tulehdusta edistävien rasvahappojen pitoisuudet verenkierrossa olivat kohonneet. Endometrioosikudoksilla tehdyssä työssä löysimme joukon munasarjan endometrioosille ominaisia proteiineja, joista monet olivat kasvutekijä- TGF-β1-säädeltyjä. Tulokset luovat pohjaa jatkotutkimuksille endometrioosin diagnostiikan kehittämiseksi.

Avainsanat: Kudos, massaspektrometria, proteomiikka, biomarkkeri, merkkiaine, metabolomiikka, transkriptomiikka, maksa, aromataasi P450, hiiri, endometrioosi, munasarjan endometrioosi, tilastotiede, Reproducibility-optimized test statistic (ROTS)

CONTENTS

ABBREVIATIONS

2-DE	Two-dimensional gel electrophoresis
ACN	Acetonitrile
AIMS	Accurate inclusion mass screening
ArKO	Aromatase knockout mice
AROM+	Aromatase overexpressing mouse model
AUC	Area under curve
CID	Collision induced dissociation
CPLL	Combinatorial peptide ligand libraries
CV	Coefficient of variance
DDA	Data dependent analysis
DIGE	Difference gel electrophoresis
ECD	Electron capture dissociation
EMT	Epithelial mesenchymal transition
ESI	Electrospray ionization
ETD	Electron transfer dissociation
FDR	False discovery rate
GnRH	Gonadotropin-releasing hormone
HCD	Higher energy collisional dissociation
HPLC	High performance liquid chromatography
IPA	Ingenuity pathway analysis
iTRAQ	Isobaric tags for absolute and relative quantification
LC	Liquid chromatography
LIMMA	Linear Models for Microarray Data
LIT	Linear ion trap
LOD	Limit of detection
LOQ	Limit of quantification
m/z	Mass-to-charge ratio
MALDI	Matrix assisted laser desorption ionization
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MS1	Precursor ion scan
MS^E	All-ion fragmentation

mTRAQ	Non-isobaric tags for absolute and relative quantification
PCA	Principal component analysis
PE	Patient endometrium
PO	Ovarian endometrioma
PSM	Peptide-spectrum match
PTM	Post translational modification
Q	Quadrupole
qRT-PCR	Quantitative real-time reverse transcriptase - polymerase chain reaction
RIA	Radioimmunoassay
ROC	Receiver operating characteristic
ROTS	Reproducibility-optimized test statistic
RP	Rank product
RT	Retention time
SAM	Significance Analysis of Microarrays
SAX	Strong anion exchange
SCX	Strong cation exchange
SDS	Sodium dodecyl sulfate
SELDI	Surface-enhanced laser desorption/ionization
SILAC	Stable isotope labeling with amino acids in cell culture
SISCAPA	Specific antipeptide antibody
SpC	Spectral count
SRM	Selected reaction monitoring
SWATH	Sequential window acquisition of all theoretical fragment ion spectra
TGF-β1	Transforming growth factor beta-1
TMT	Tandem mass tag
TOF	Time-of-flight
UPLC	Ultra-high performance LC
WT	Wild type
XIC	Extracted ion chromatogram

LIST OF ORIGINAL PUBLICATIONS

The thesis is based on the following original publications, referred to in the text by Roman numerals **I-III**.

I. Anna Pursiheimo*, **Anni P. Vehmas***, Saira Afzal*, Tomi Suomi, Thaman Chand, Leena Strauss, Matti Poutanen, Anne Rokka, Garry L. Corthals, and Laura L. Elo. (2015). Optimization of Statistical Methods Impact on Quantitative Proteomics Data. *Journal of Proteome Researc*h 14: 4118-4126.

II. **Anni P. Vehmas**, Marion Adam, Teemu D. Laajala, Gabi Kastenmüller, Cornelia Prehn, Jan Rozman, Claes Ohlsson, Helmut Fuchs, Martin Hrabě de Angelis, Valérie Gailus-Durner, Laura L. Elo, Tero Aittokallio, Jerzy Adamski, Garry Corthals, Matti Poutanen, and Leena Strauss. (2016). Liver lipid metabolism is altered by increased circulating estrogen to androgen ratio in male mouse. *Journal of Proteomics* 133: 66-75.

III. **Anni P. Vehmas**, Dorota Muth-Pawlak, Kaisa Huhtinen, Taija Saloniemi-Heinonen, Kimmo Jaakkola, Teemu D. Laajala, Heidi Kaprio, Pia A. Suvitie, Tero Aittokallio, Harri Siitari, Antti Perheentupa, Matti Poutanen, and Garry L. Corthals. (2014). Ovarian Endometriosis Signatures Established through Discovery and Directed Mass Spectrometry Analysis. *Journal of Proteome Researc*h 13: 4983-4994.

* Equal contribution

In addition, some unpublished data is presented. The original publications are reproduced with the permission of the copyright holders.

1. INTRODUCTION

As is the case with any biological system, a tissue has its structure and dynamics, which determine its physical properties and functions within the body. These tissue characteristics are modulated by gene expression, protein interactions and biochemical pathway activity, which generate complex networks that cannot be understood by studying singular molecules, one at a time. To gain a better understanding, comprehensive datasets of gene, protein and metabolite expression are studied with current omics technologies.

Contemporary estimates of the size of the human genome lie at about 20 300 protein-encoding genes. However, the actual proteome size is a much larger number due to nonsynonymous single nucleotide polymorphisms, alternative promoters and transcription start sites, alternative splicing patterns and post translational modifications (PTMs) (Harrow et al, 2012). The method of choice for the proteome exploration is mass spectrometry (MS), where very complex mixtures of proteins are first digested to peptides with an endoprotease. The peptide fragments are then separated by one or more dimensions of liquid chromatography (LC) to reduce the complexity of the sample, after which the masses of the peptides are measured in a gas phase by MS.

A mass spectrometer consists of an ion source, where the charged peptides enter the mass spectrometer, a mass analyzer that measures the mass-to-charge-ratio (m/z) of the analytes, and a detector that records the number of ions at each m/z, constructing the mass spectrum. MS based proteomics allows multiplexed data generation, i.e., the analysis of multiple proteins within one sample during one analysis, and challenges antibody based methods such as ELISA and Western blot (Woods et al, 2014) in protein analytics. It is possible to routinely identify 5000 proteins within one MS run (Michalski et al, 2011b; Thakur et al, 2011) and the analysis of proteomes consisting of 10 000 -20 000 identified proteins have been reported (Kim et al, 2014; Wilhelm et al, 2014).

In quantitative proteomics, a myriad of different techniques exist, suitable for different types of applications and sample matrices. Label-free quantitative methods have emerged to challenge label-based quantitative methods in MS-based proteomics. These label-free approaches use the signals obtained directly from the peptides undergoing MS analysis to acquire a quantitative measure. When compared to methods using labels, the label-free approaches are inexpensive, flexible and scalable to hundreds of

samples, which make them advantageous in large biomarker discovery projects (Sandin et al, 2015). When a verified set of quantified proteins is combined with different levels of molecular data, such as gene expression or metabolite information, the downstream effects can be studied on a system-wide scale. Even though the knowledge of such systems is still limited, the understanding of molecular networks achieved with different omics technologies is now comprehensive enough to monitor, on a tissue level, the disturbances caused by a transgene, disease or medication, for example.

2. REVIEW OF THE LITERATURE

2.1 Protein and peptide identification

In a typical bottom-up shotgun liquid chromatography – tandem mass spectrometry (LC-MS/MS) experiment, mixtures of proteins are first subjected to endoprotease digestion. The resulting peptides are subsequently separated by liquid chromatography (LC) and upon elution brought to gas phase during electrospray ionization, and finally mass analyzed by mass spectrometry (MS). The MS instrument is operated in a data dependent manner (DDA), where abundant precursor ions are sequentially isolated by MS and fragmented by tandem MS (MS/MS). The combination of MS and MS/MS data can then be used together by database search algorithms, such as Sequest, to identify the peptide sequence of each precursor/product ion series. The subsequent protein identification is then accomplished based on its identified peptides (Eng et al, 1994).

2.1.1 LC/ESI-MS/MS instrumentation

In practice, peptides can be charged and brought into the gas phase by popular ionization methods known as MALDI (matrix assisted laser desorption ionization; Hillenkamp & Karas, 1990) and electrospray ionization (ESI; Fenn et al, 1989). Both MALDI and ESI are commonly used, however, even though SELDI (surface enhanced laser desorption/ionization; Hutchens & Yip, 1993) has been used in some biomarker discovery projects, inline LC/ESI coupled to MS/MS is the method of choice in most proteomics workflows globally.

Hybrid mass spectrometers are the most popular MS systems used in proteomics today. They consist of MS configurations with different ion optics, mass analyzers and fragmentation sources. A quadrupole mass analyzer (Q) is essentially a molecular mass filter with alternating electric potentials that only allow ions with a certain m/z to pass through. In linear ion trap (LIT), both isolation and fragmentation of peptide ions is possible by sequential stabilization, fragmentation and ion ejection events within the trap. Time-of-flight (TOF) analyzers measure the mass of an ion by measuring its mass-dependent time of flight from an ion source to the detector. These analyzers are extremely fast, have a high mass range and resolving power up to 60 000 (mass / full peak width at half maximum). In Orbitrap, ion currents of peptides are detected as

they oscillate an orbital electrode. The ion currents are converted to m/z values, based on their frequencies by the use of Fourier transformation providing a very high mass resolving power of up to 100 000 (Michalski et al, 2012; Scigelova & Makarov, 2006; Woods et al, 2014).

Multiple peptide fragmentation methods exist, each providing complementary peptide patterns suitable for different applications. The most common fragmentation methods used are collision induced dissociation (CID) and higher energy collisional dissociation (HCD). CID and HCD predominantly generate b- and y-type ions (Figure 1) by disrupting the peptide backbone between the C- and N-terminal amino acid residues producing positively charged peptide fragments. HCD provides high intensity reporter ion spectra that enables reliable identification of precursor ions (Ting et al, 2009). Two additional methods, electron capture dissociation (ECD; Zubarev et al, 1998) and electron transfer dissociation (ETD; Syka et al, 2004) generate mostly c- and z-type of ions (Figure 1). Of these, especially ETD is considered as a soft ionization technique suitable for the analysis of phosphopeptides, since it retains the phospho-groups in peptide backbone (Sarbu et al, 2014). Also the successful combination of CID and ETD fragmentation has been demonstrated while analyzing glycosylation and complex proteomes (Hanisch, 2012; McAlister et al, 2008; McAlister et al, 2007).

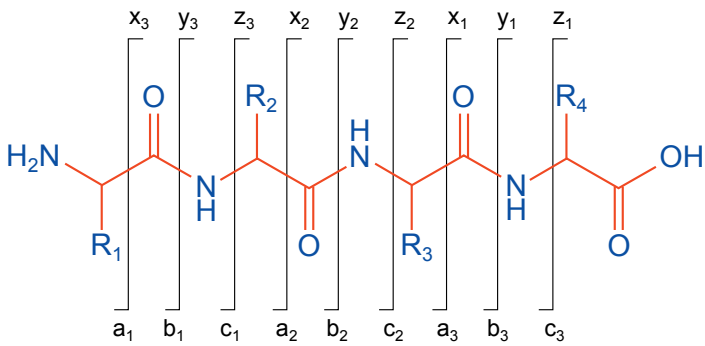

Figure 1. The common nomenclature of peptide fragmentation by Roepstorff and Fohlman (1984). Different fragmentation techniques disrupt the peptide backbone in different positions generating fragment ions, termed a, b and c -ions (N-terminal) or x, y or z -ions (C-terminal). Modified from Roepstorff and Fohlman (1984).

The development of hybrid instruments has improved the use and performance of mass spectrometers that can now offer a combination of high-quality MS data on different key characteristics, such as improved mass accuracy, resolution, scan rate, dynamic

range, and overall peptide coverage in bottom-up workflows. Previously, the high speed peptide identification, where the ions are gathered, isolated and fragmented by CID inside the trap, made the LIT system one of the most used mass analyzers in proteomics. However, when high resolution and mass accuracy in precursor analysis and higher scan speeds were required, such as in LC-MS label-free quantification (see 2.2.2), ion traps have successfully been used in combination with the Orbitrap analyzers (Makarov et al, 2006; Yates et al, 2006). Currently Q-hybrid instruments are frequently used; alternatives for peptide analysis are the Q-TOF and Q-Orbitrap (Q-Exactive) setups. With the Q-TOF instrument it is possible to filter and fragment precursor masses sequentially, which is important for peptide identification in shotgun proteomics experiments. Q-TOF systems have high mass accuracy combined with millisecond scan rates (Beck et al, 2015). The Q-Orbitrap instrumental setup functions similarly to Q-TOF, only difference being that HCD fragmentation is performed and mass analysis is made in Orbitrap with slightly reduced scan rates. A high-resolution QQ-TOF (Triple-TOF) instrument enables both DDA and data independent fragmentation of all ions in extremely wide precursor isolation windows, which produces time-resolved MS/MS spectra (Schilling et al, 2015). It has been shown that with this so called SWATH method peptide quantification and identification can be performed in a concentration range of up to four orders of magnitude, sensitivities ranging from 0.6 fmol to 1.3 pmol (Gillet et al, 2012). Furthermore, recently a Q/LIT-Orbitrap (Orbitrap Fusion) instrument demonstrated scan rates of 20 cycles per second (20 Hz) (Hebert et al, 2014; Senko et al, 2013).

2.1.2 Generic shotgun proteomics workflow

Proteome analysis technologies are applied for identification, quantification and characterization of proteins in various biological systems. Despite its resemblance to genomics, proteomics encounters more challenges than genome sequencing. This is partly because of the immense complexity of the proteome; proteins are expressed in differential tissues and cellular locations, often transiently, and have multiple interactions and post-translational modifications. A shotgun proteomics experiment aims at accurate identification and contextualization of as many proteins in the sample matrix as possible. Currently in shotgun proteomics, complex mixtures from few hundred proteins to nearly 20 000 proteins are analyzed (Kim et al, 2014; Wilhelm et al, 2014). The proteins can be extracted from prokaryote or eukaryote cells, tissues or body fluids and subjected to direct LC-MS/MS analysis, or analyzed

after fractionation and/or affinity-based purification. As comprehensive LC-MS/MS procedures in current use are more geared towards low molecular weight substances, the proteins are digested to peptides prior to the MS analysis. This is usually accomplished by applying an amino acid sequence specific protease, such as trypsin or Lys-C (Gershon, 2014).

For the analysis of complex mixtures of peptides, reversed phase high performance LC (HPLC) is the preferred in-line separation method with one to several hours long aqueous/organic gradient, and with nanoliter/min flow rates. However, often an additional separation step is added to improve the detection of co-eluting peptides, to increase the proteome coverage or to decrease the effect of the complex background. The samples can be processed to fractions by electrophoretic approaches such as one (1D SDS-PAGE) or two-dimensional gel electrophoresis (2-DE), where the proteins are separated according to their molecular weight and isoelectric point, by various liquid chromatographic methods or by depleting highly abundant proteins from the sample using affinity methods. Increasingly popular choices for sample pre-processing are techniques applying strong cation exchange chromatography (SCX) for the enrichment of acetylated and phosphorylated peptides (Dephoure et al, 2008; Taouatas et al, 2009), or strong anion exchange (SAX; Wiśniewski et al 2010) or reversed phase chromatography at high-pH (Delmotte et al, 2007). Also hydrophilic interaction chromatography has been found to simultaneously increase proteome coverage and phosphopeptide enrichment (Di Palma et al, 2011; McNulty & Annan, 2008). Currently the trend is to use long LC-columns exceeding the length of 20 cm packed with small particles (<3 µm) together with long gradients of 2 hours or more to improve peptide separation (Eeltink et al, 2009; Yamana et al, 2013). However, the increase of back pressure by the use of long columns and small particles require the use of LC systems capable of handling these pressures such as ultra-high performance LC (UPLC). Alternatively, the back pressure can be reduced by heating the column to 40-60 ˚C (Hyung et al, 2011; Motoyama et al, 2006; Nagaraj et al, 2012).

By these enhancements higher peak capacity can be reached, determined by the number of distinct peaks eluted over the chromatographic gradient. Chromatographic peaks as narrow as 4-10 seconds (full width at half maximum) can be gained, which increases the peak resolution and sensitivity of the system due to more concentrated analytes (Köcher et al, 2011b). Mass spectrometers are constantly evolving towards faster and more sensitive systems having higher mass accuracy and resolution,

resulting to devices collecting MS-spectra routinely at a high resolving power, which lead to more confidence in protein identification. The LC-MS/MS setups can reach attomole range sensitivity, detecting over 100 000 peaks per hour and sequencing up to 20 peptides per second also in complex sample matrices (Gillet et al, 2012; Hebert et al, 2014; Pelander et al, 2011). Despite the improvements, in most samples the molecular complexity and dynamic range of protein concentrations are still beyond the capacity of current instruments (Gillet et al, 2012; Michalski et al, 2011a; Richards et al, 2015). Therefore, data independent approaches, such as MS^E (Silva et al, 2005) and SWATH (Gillet et al, 2012), where all ions at certain mass range extending from 2.5 m/z to full m/z range are fragmented simultaneously, are gaining interest (Chapman et al, 2014).

2.1.3 Peptide identification

During the constant elution of the sample in LC and ensuing ionization of peptides, the mass spectrometer scans the peptide masses over a mass range from 300-2000 Da in approximately 1-2 seconds. Subsequently, the instrument, guided by acquisition software, selects and isolates number of precursor ions for sequential fragmentation and then returns to scanning the peptide masses present. The time the instrument uses for each MS precursor scan and the subsequent MS/MS fragmentation is often referred as the duty cycle of the instrument. The number of fragmented precursors can be fixed for every cycle or the instrument can operate in DDA-mode, where only precursors exceeding a certain signal threshold are chosen for fragmentation. The fragmentation spectrum of each precursor is called tandem or MS/MS spectrum; in current instruments typically 5-20 MS/MS product spectra are produced for each MS scan.

The precursor peptide mass is always associated with the fragment masses and together this information is used for peptide identification. In data independent approaches peptides are identified by the alignment of the peptide and fragment ions based on their exact matching retention times (Bern et al, 2010). The spectra acquired from MS and MS/MS are converted to peak lists and compared to amino acid or genomic sequence databases by an algorithm that queries the best corresponding theoretical spectra (Hoopmann & Moritz, 2013). There are three types of search algorithms: 1) sequence-based search engines such as X!Tandem (Craig & Beavis, 2004), Mascot (Perkins et al, 1999), Sequest (Eng et al, 1994) and Comet (Eng et al, 2013) that compare the acquired

spectra with theoretical spectra; 2) spectral library based engines such as SpectraST (Lam et al, 2007), X!Hunter (Craig et al, 2006) and MzMod (Horlacher et al, 2015) that compare the acquired spectra with spectra in a spectral library and 3) de novo search engines such as PEAKS (Ma et al, 2003), PepNovo (Frank & Pevzner, 2005) and UniNovo (Jeong et al, 2013), calculating a peak pattern based on the acquired spectra without previous knowledge of a sequence.

The output of database search algorithms is a list of peptide-spectrum matches (PSMs) each associated with a p-value or a score determining the quality of the match. However, the score determines only the quality of a single peptide-spectrum match, but not the amount of incorrect, false positive PSMs in the complete dataset. Therefore, a statistical correction for multiple testing is required. In proteomics, the estimate of false positive identifications is traditionally performed by using Benjamini-Hochberg false discovery rate (FDR) method. The method is based on the measurement of the ratio of incorrect PSMs to all accepted PSMs. To estimate the number of incorrectly identified PSMs in the dataset, the spectral files are searched against a decoy database that contains shuffled, randomized or reverse versions of all the sequences in the original database. When searched together with the real database, the PSMs identified in decoy database can be considered false positives and used for FDR calculation (Choi & Nesvizhskii, 2008; Käll et al, 2008). For reliable protein identification, typically two identified peptides are required. Advances in the determination of false positives (Keich et al, 2015; Savitski et al, 2015), quality requirements of protein identity (Martínez-Bartolomé et al, 2013; Taylor et al, 2008) and high resolution mass spectrometers available as well as improved computational tools now enable reliable identification of more than half of all MS/MS spectra generated in an experiment. Commonly used search engines have as an option to specify few peptide modifications or mass shifts e.g. methionine oxidation often occurring during sample handling. However, not all possible modifications can be chosen without increasing the search space and compromising the quality of identifications. Nevertheless, some unidentified spectra always remain in the data, even though the new hybrid instruments with high scan speeds have improved the situation (Chick et al, 2015).

2.2 Protein quantification in complex samples

Multiple quantification methods are applied to clinical samples in mass spectrometry based proteomics. Quantitative approaches include electrophoresis based methods,

such as difference gel electrophoresis (DIGE; Unlü et al, 1997), where the quantification is accomplished by the labeling of proteins with cyanine dyes, or protein microarray-based approaches such as SELDI (Hutchens & Yip, 1993). However, these methods are not as comprehensive in proteome screening as MS-based quantitative strategies, which are discussed more thoroughly in this chapter and introduced in Figure 2. The quantitative MS-based approaches can be divided roughly to stable isotope labeling based methods, label-free methods and to targeted methods (Figure2). Stable isotope labeling based methods employ labeling strategies, such as iTRAQ (Ross et al, 2004), that are useful only with very small cohorts (<10). The label-free and targeted methods, such as selected reaction monitoring, are more suitable for larger sample numbers, up to 100 samples (Gallien et al, 2011).

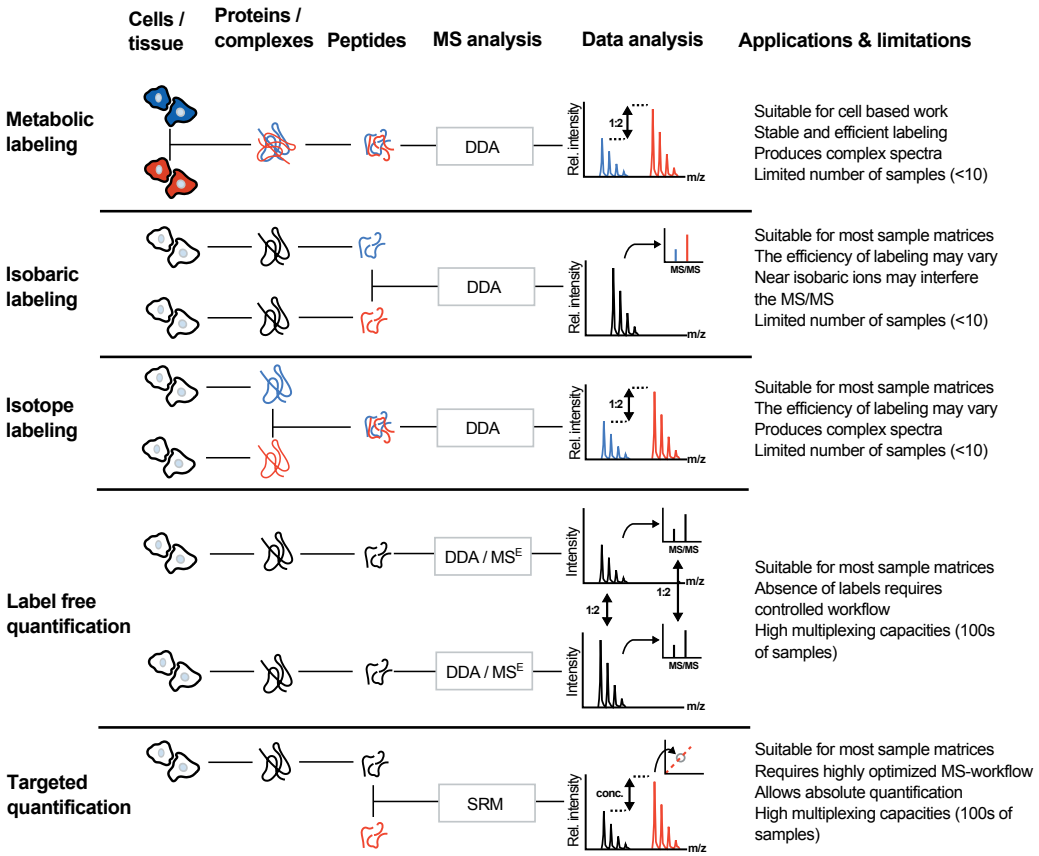

Figure 2. An outline of the MS-based quantitative approaches in proteomics. DDA, Data dependent analysis; MS^E, All-ion fragmentation; SRM, Selected reaction monitoring. Modified from Thermo Fisher Scientific.

2.2.1 Methods based on stable isotope labeling

Stable isotope based quantitative methods are applied in proteomics when a small number of comparisons are sought. As the physicochemical properties of stable-isotope labeled peptides and natural peptides are similar, they also demonstrate in very similar behavior in MS. Nonetheless, the isotope labeled peptide can be distinguished from its natural version based on a mass shift specific for the heavy label (Figure 2). Thus, the equivalent physicochemical properties of the labeled peptide with the natural version allow relative and absolute quantification of a peptide or a protein by comparing the intensities of the labeled and natural version of the analyte.

Metabolic labeling of bacteria and yeast can be accomplished by culturing them in a presence of ^{15}N resulting in peptides with heavy nitrogen atoms only (Gouw et al, 2010; Oda et al, 1999). The disadvantage of the method is, however, that it generates complex spectra, which can be challenging to interpret as all nitrogen atoms in each peptide are isotopically labeled. In another metabolic labeling technique, SILAC (stable isotope labeling with amino acids in cell culture) method (Ong et al, 2002), isotope-labeled lysine and arginine (^{13}C, ^{15}N) are incorporated into protein from the cell culture medium. In this way, a labeled standard for each protein is generated. Originally, SILAC analysis was developed for three samples unless deuterated amino acids were used. The use of deuterated amino acids, however, affect the retention times of peptides and cause overlapping isotope clusters, which complicates data analysis (Zhang et al, 2001). Yet, more recently the method has been used in combination with MS1 based label-free quantification to extend the quantitative capacity to five samples (Molina et al, 2009). Other recent developments have been the pulsed SILAC experiments to monitor protein turnover and response rates in cells (Cambridge et al, 2011; Milner et al, 2006; Schwanhäusser et al, 2009), absolute quantification by SILAC (Hanke et al, 2008) and the super SILAC method. The super SILAC application has been adapted to quantify proteins from tissue using a mixture of SILAC-labeled cell lines as an internal standard for quantification (Geiger et al, 2010). More recently, the entire proteome of *C. elegans* was successfully labeled with ^{15}N-lysine (Fredens & Færgeman, 2012).

Several methods for isobaric labeling exist. Here the N-terminus and ε-amino group of lysine in peptides or proteins are labeled with stable-isotope labels. The principle of isobaric labeling is that peptides labeled with different tags have isobaric masses

in MS, but altered fragmentation pattern in MS/MS (Figure 2). Each individual tag can be distinguished and the sample quantified based on the intensity of a specific low molecular mass reporter ion. Currently popular methods, TMT (Tandem mass tag), iTRAQ (isobaric tags for absolute and relative quantification) both label the primary amines of the amino acid backbone (Fredens & Færgeman, 2012; Ross et al, 2004; Thompson et al, 2003; Wiese et al, 2007). An advantage of the isobaric labeling when compared to SILAC strategy is that it allows multiplexing up to 10 samples without adding complexity to LC separation or peptide mass spectra (Murphy et al, 2015).

A fairly common problem in isobaric strategies is the presence of near isobaric ions that are isolated and fragmented together with the ion of interest (Karp et al, 2010). Solutions for this problem have been generated (Ting et al, 2011; Wenger et al, 2011). However, they are accompanied with a reduced sensitivity and acquisition speed. The mTRAQ (Non-isobaric tags for absolute and relative quantification) labels have been applied for MS-level quantification where the isotope labels can be distinguished in peptide mass spectra (Figure 2; DeSouza et al, 2008). More recently, when the mTRAQ approach was compared to SILAC, it was concluded that mTRAQ was capable of reliable quantification but presented slightly more variance in terms of mass ratios (Oppermann et al, 2013). From other isotopic labeling strategies, stable-isotope dimethyl labeling (Hsu et al, 2003) has gained popularity in the recent years because of its simplicity (Zhou et al, 2014). Finally, even though labeling in protein level would be possible with the isobaric methods, in practice usually peptides are labeled in order to avoid adding complexity or diluting the signal by incomplete labeling.

2.2.2 Label-free quantification methods

Label-free methods have gained popularity in large clinical projects, where high amount of samples need to be compared, as using isotope labels, the labeling of maximum of 10 samples is possible (Murphy et al, 2015). The label-free methods can be further divided into two main approaches, MS (MS1) and MS/MS (MS2) level quantification methods. MS1 level methods measure and quantify the signal intensities derived from a peptide before it is fragmented in MS/MS, whereas the MS/MS methods, more commonly spectral count (SpC) methods, measure and quantify the amount of the fragments for each peptide (Figure 2). Some methods also consider the intensities of the fragment

ions. From these two label-free approaches, a reasonably good correlation has been found. However, it has been shown that MS1 methods provide better solution for newer instruments providing both high mass accuracy and high resolution (Choi et al, 2012; Grossmann et al, 2010).

In SpC approaches the quantity of each protein is measured indirectly by counting the number of the MS/MS spectra of each precursor ion or the number of PSMs (Li et al, 2012; Washburn et al, 2001; Zhang et al, 2006a). The method is based on the DDA approach, where a set of precursors are chosen for fragmentation in each MS-MS/MS cycle (See chapters 2.1.2 and 2.1.3). As all of the database search algorithms deliver information of the number of fragmented spectra of each peptide, a spectrum count for each protein can be calculated. This number can vary from less than two for low abundant proteins, 2-10 for medium-abundance proteins and over 10 for high abundant proteins. Thus, in order to maximize the number of counts for low abundant proteins in SpC, the MS method should be optimized for this quantification method. By the use of dynamic exclusion, previously fragmented ions are rejected for a predefined time inside the MS. The dynamic exclusion time needs to be optimized to prevent the repeated fragmentation of abundant precursor ions with wide peak widths masking low abundant peptides. In contrast, the low abundant peptides typically have narrow peak widths and short elution times, which require fast scan speeds from the instrument. Due to these and some other reasons the quantification of peptide mixtures with high dynamic range of peptide concentrations and very small proteins may be challenging (Zhou et al, 2012).

Dynamic range of quantification can be improved in SpC by including peptides with lower confidence PSMs. Stricter criteria for PSMs, however, has shown to improve statistical significance of smaller changes (Cooper et al, 2010). The protein inference problem, i.e., the same peptide sequence being present in multiple proteins (Nesvizhskii & Aebersold, 2005), has been studied in label-free quantification (Zhang et al, 2010). The results indicate that the best solution of the problem can be found by distributing the shared spectrum counts between different proteins proportionally to the number of unique spectrum counts measured for each protein (Zhang et al, 2010). Additionally, as in any label-free quantification method, in SpC, the LC-MS/MS operation should be stable and all samples analyzed in a single batch. Indeed, when SpC approach was compared to isotope labeling methods using the LTQ Orbitrap Velos system, it was concluded that SpC provided the most comprehensive quantitative coverage of the

sample, but provided less precise results due to lower repeatability (Li et al, 2012). The variations may originate from sample loading, chromatography and the data dependent precursor fragmentation setup. However, it has been suggested that these variations can be adjusted by normalizing the data for the number of PSMs in each sample (Old et al, 2005). Also when the spectrum count was divided by the average total-ion count in a corresponding spectrum, a clear improvement in accuracy and linear dynamic range was detected (Asara et al, 2008). The normalization including intensity information has been later on improved by a specific protein spectral index method (Colaert et al, 2011; Griffin et al, 2010).

MS1 methods measure ionized peptides directly and avoid the stochastic sampling of MS/MS method, where only the most intensive precursor ions are fragmented and quantified. Moreover, it has been shown that the intensity of ESI is linearly correlated with peptide abundance in over four orders of magnitude, which allows reliable peptide quantification based on its MS1 signal (Bondarenko et al, 2002). In the MS1 approach, the MS spectra are used to generate extracted ion chromatograms (XICs) in which the peptide signals, features, are defined by their m/z, retention time and intensity. The abundance of each peptide feature in the dataset is then calculated by the use of this data. Also, multiple features can denote each peptide sequence presenting its charge states. The peptide ions representative to each feature can be chosen for fragmentation in MS/MS and used for identification. Therefore, a critical step of MS1 intensity based label-free quantification is the correct assignment of peptide features to the right sequence. Thus, the LC-MS configuration needs to have high enough resolving power and accuracy to avoid co-elution of peptides and to ensure right precursor-sequence match (Sandin et al, 2015; Sun et al, 2005). In order to reach biological conclusions of the results, the peptide sequences are matched with corresponding proteins. This is often accomplished by grouping the peptide signals into protein identifiers by either calculating the average or sum of peptide features of a protein (Clough et al, 2009).

To achieve reliable quantification in label-free MS1 analysis, samples need to be prepared reproducibly (Ong & Mann, 2005) and have 8-10 measured spectra for each peptide ion to reach good peak shape over differential ion concentrations (Radulovic et al, 2004). As the number of achievable spectra depends on the scan time for measured ions, a robust analysis scheme for MS1 quantification requires specific considerations of the LC-MS workflow. Faster MS1 scan time and duty cycle produce a higher number

of MS1 signals with lower amount of spectra per run resulting in more peptide identifications but compromised quantitative quality. Therefore, the optimal MS1 set-up is often a compromise between reliable quantification and amount of identified peptides. Approaches have been developed, where the data acquisition of MS and MS/MS are performed separately to ensure the best possible quality of quantification data (Fang et al, 2006). These approaches require specific tools to match the quantitative peptide feature and its sequence, but have shown to be advantageous in biomarker studies (Varnum et al, 2011). Another methodology well suited together with MS-level quantification is the all-ion fragmentation or MS^E that has recently proven to increase coverage when compared with traditional Q-TOF workflow (See section 2.1.1, Blackburn et al, 2010). Moreover, the SWATH-method employs fragment ion intensities in the quantification of the corresponding peptide signals (See section 2.1.1, Gillet et al, 2012).

As with other quantification methods, the experimental variation should be monitored in label-free approaches (Bondarenko et al, 2002), and sufficient number of peptides per protein identified to avoid missing values in the data matrix (Ning et al, 2012). A challenge is also to extend the dynamic range of the analysis to include the measurement of low abundant proteins (Zubarev, 2013). However, good signal-to-noise ratios allow peptide quantification in a higher dynamic range and can be reached by narrowing the LC peak width by UPLC or long column HPLC (Sandra et al, 2009; Sandra et al, 2008). Retention time shifts can also affect peptide co-elution, which in turn can have an effect on ionization efficiency, as different peptides compete for ionization under different chromatographic conditions (Sun et al, 2005).

The analysis workflow for MS1 quantification usually starts with raw spectral data and consists of mass calibration, noise and data reduction, after which the alignment of the feature maps in the dimension of time of all the analyzed samples is performed. The workflow continues by the definition of the corresponding features in each sample and proceeds with data normalization. In the normalized map the relative quantification of features across all samples in an experiment is possible (Nahnsen et al, 2013). There are many software platforms designed for the crucial steps of data analysis such as alignment, feature finding and normalization as listed in Table 1.

Table 1. A list of software packages used for LC-MS quantification. Drawn according to Nahnsen et al (2013) and respective publications.

Name	Input formats	Resolution	Quantification	Statistical analysis	Reference
Census	mzXML, MS1, MS2, pepXML, DTASelect	Low and high	SpC, MS1	-	(Park et al, 2008)
Corra	mzXML, pepXML	Low and high	MS1	+	(Brusniak et al, 2008)
Expressionist	mzXML, NetCDF, major vendors	Low and high	MS1	+	Genedata
IDEAL-Q	mzXML	Low and high	MS1	-	(Tsou et al, 2010)
MAPA/ ProtMAX	mzXML	High	SpC, MS1	-	(Egelhofer et al, 2013)
MapQuant	Thermo .RAW	Low and high	MS1	-	(Leptos et al, 2006)
Mascot Distiller	mz(ML|XML), major vendors	Low and high	SpC, MS1	-	Matrix Science
MaxQuant	Thermo .RAW	High	MS1	+	(Cox & Mann, 2008)
msInspect	mzXML, mzML	Low and high	MS1	-	(Bellew et al, 2006)
mzMine 2	mz(ML|XML|Data), Thermo .RAW, NetCDF	Low and high	MS1	+	(Pluskal et al, 2010)
OpenMS/ TOPP	mz(ML|XML|Data)	Low and high	MS1	-	(Kohlbacher et al, 2007; Sturm et al, 2008)
Progenesis LC-MS	mz(ML|XML), major vendors	Low and high	MS1	+	Nonlinear Dynamics
ProteinQuant Suite	mz(ML|XML|Data)	Low and high	MS1	-	(Mann et al, 2008)
ProteoIQ	mz(ML|XML|Data), major search engines	Low and high	SpC, MS1	+	Premier Biosoft
Proteios	mz(ML|XML|Data)	Low and high	MS1	-	(Sandin et al, 2013)
pView 2	mzXML, pepXML	High	MS1	+	(Khan et al, 2009)
RIBAR/xRIBAR	ms_lims, Mascot .dat	-	SpC	-	(Colaert et al, 2011)
Scaffold	major search engines	-	SpC	+	Proteome Software
SIEVE	Thermo. RAW	Low and high	MS1	+	Thermo Scientific
Spectrolyzer	mz(ML|XML|Data), major vendors	Low and high	MS1	+	Medic Wave
SuperHirn	mzXML, pepXML	High	MS1	-	(Mueller et al, 2007)
Viper	PEK, CSV, mz(XML|Data)	High	MS1	-	(Monroe et al, 2007)

2.2.3 Targeted quantification methods

In proteomics, the selected reaction monitoring (SRM) is a stable isotope based method suitable for multiplexing peptide measurements. SRM is a targeted method that screen predetermined precursor masses and their fragmentation products to detect and

quantify them using stable isotope-labeled synthetic peptides (Figure 2). A key term in a SRM experiment is “transition”, which refers to each targeted precursor ion and its fragmentation products. Triple-quadrupole instruments are used in SRM, where the first quadrupole acts as a mass filter for the precursor ion, the second quadrupole is used for CID of the ion and the third quadrupole measures the produced fragment ions. Because of this setup, every transition represents an independent assay, which enables robust and highly sensitive quantification of peptides in a dynamic range up to five magnitudes in a complex sample matrix (Gallien et al, 2011).

When building a SRM method, the peptides picked need to be unique for the protein of interest. Preferably peptides need to be easily identified in MS, and have as few modification sites or missed cleavages as possible (Duncan et al, 2009). The fragment ions used for quantification are chosen from MS/MS spectra that can be acquired experimentally or originate from a public database. For this purpose databases such as PeptideAtlas (Desiere et al, 2006) and SRMAtlas that contains previously acquired spectra (Picotti et al, 2008), are often used. After the determination of the precursor-fragment ion pairs, the transitions, the optimal instrument parameters are determined for each transition. The transitions are also tested for selectivity. The selectivity of a transition depends heavily on a sample matrix and, thus, need to be tested for every new sample matrix type. Typically fragment ions from co-eluting close-isobaric ions can have a large effect on selectivity of the transition (MacCoss et al, 2003). After this, the lower limits of detection and quantification are determined. Finally the complete method is compiled including all transitions with a precursor mass, its fragment ion masses, elution time, collision energy and dwell time, which can be defined as the time spent acquiring a specific transition during each cycle. In designing the SRM method, one major obstacle is the confirmation of peptide identity. This can be performed based on the retention time or more precisely by MS/MS (Parker et al, 2014).

Due to the successful application of SRM MS, multiple tools and approaches have been developed recently. When a large scale SRM experiments are designed, many features are needed for multiplexing, such as number of transitions for each peptide, dwell time and the time between each transition. Now, by the use of these techniques, the quantification and identification of hundreds of peptides in one experiment is possible. Additional improvements especially in SRM selectivity can be reached by using an instrument capable of ion mobility separation of isobaric precursor ions (Klaassen et

al, 2009) and sensitivity in SRM^3, if a linear ion trap instrument is used for a second fragmentation step of the fragment ion (Fortin et al, 2009; Jeudy et al, 2014). In "intelligent SRM" eight to ten additional transitions, covering whole fragment ion series of a precursor, are triggered in a data dependent manner. The precursor ion of the artificial MS/MS spectra is then identified by spectral matching using spectral libraries (Kiyonami et al, 2011). Moreover, parallel reaction monitoring (PRM), a method allowing high resolution acquisition of MS/MS spectra with minimum optimization steps, has been developed using high resolution quadrupole instruments, such as QExactive (Dillen et al, 2012; Peterson et al, 2012).

2.3 Systems-wide proteomics for disease interrogation

2.3.1 Biomarker workflow using mass spectrometry

Biomarkers are developed to estimate a risk to develop a disease (Schrag et al, 2000), for early detection of disease (Etzioni et al, 2003), for disease classification (Yu & Hung, 2000) or to monitor the success of the treatment (Hughes et al, 2006). Even though single protein biomarkers are currently mostly used at the clinics, such as troponin for heart attack and human chorionic gonadotropin for pregnancy, novel disease biomarkers are expected to be composed of a panel of proteins (Frantzi et al, 2014). Depending on the use of the biomarker panel, it needs to fulfil different criteria: For example, a biomarker designed for early cancer detection in millions of people needs to show extremely high specificity to avoid a large number of false positive diagnoses, while the specificity can be lower for a biomarker used to monitor a treatment response of an identified disease, because the number of patients is lower. In order to be of benefit, the test must be directed towards a disease with treatment options that can prolong or improve the quality of life of the patient. An example of successful development of a biomarker panel using proteomics is the OVA1 test for ovarian cancer. This test consists of five protein biomarkers and is used together with clinical assessment to predict the likelihood of malignancy (Longoria et al, 2014; Zhang et al, 2004).

A set of parameters can be used for the assessment of the analytical performance of an assay. The random error of any method can be evaluated with means of precision. Precision is often reported using the concepts of repeatability, intermediate precision and reproducibility. Repeatability is the variance of a controlled experiment, where

all possible parameters, such as laboratory, instrument operator, day and sample, are retained identical. This measure can be used when the instrumental precision during a sample set is evaluated. Reproducibility delineates an experiment, where such parameters vary; the experiments could be performed in different days, by a different laboratory and instrument, for example. Intermediate precision, instead, is referred as reproducibility within a single laboratory. This can be exemplified by daily analyses of a standard sample for the estimation of instrumental performance that day. Precision can be expressed as standard deviation, variance, or coefficient of variance (CV). Especially in quantitative proteomics, the precision is influenced by the analyte concentration; an assay has higher CV values for low abundant proteins and is, thus, less precise for these analytes (Andreasson et al, 2015). The systematic error of an assay is evaluated by trueness. The quantitative measure of trueness is bias, which designates the systematic difference between the experimental value and the reference value. To avoid bias, careful instrumental calibration is performed (Yau et al, 2015).

Limit of detection is the minimum amount of analyte that can be reliably identified being present in the sample. Often, with current MS setups, limits of detection (LOD) at low attomole range are reported (Thakur et al, 2011; Zhou et al, 2015). However, even if an assay is presenting low LOD for the analyte, it still may suffer from undesirable high variance and bias at low levels. For this reason the measure often reported in the context of MS based assays is the limit of quantification (LOQ). This value incorporates the effect of uncertainties and describes the concentration where the analyte can be reliably identified in the sample with acceptable bias and variance (Armbruster & Pry, 2008). Analytical interference, caused by a constituent of sample, can be defined as an effect that alters the correct reading of an analyte. In LC-MS/MS based proteomics, the largest cause of interference are the matrix effects of the ESI process, where different types of molecules are competing for ionization. These co-eluting compounds in a complex sample can cause saturation and signal suppression and have a significant effect on the accuracy of quantification (Annesley, 2003; Wilm, 2011). The effect of ionisation suppression is difficult to predict and may cause both intra and inter -sample bias. Therefore, to be considered reliably quantified, several peptides per protein are usually required and a sample set containing biological replicate samples is of importance (Wilm, 2011).

The workflow for biomarker development by MS follows a series of phases (Rifai et al, 2006), where the number of screened samples rises and the number of proteins of interest declines in each step. The phases discussed in this thesis are 1) Discovery, 2) Qualification, 3) Verification and 4) Clinical validation (Figure 3).

In the recent years, many groups have applied state-of-the-art proteomics methods to discover biomarkers for various diseases. Addona et al (2011) used discovery LC-MS/MS, accurate inclusion mass screening (AIMS), SRM and ELISAs in search of biomarkers to myocardial infarction. A similar pipeline was established for a murine model of breast cancer (Whiteaker et al, 2011). A new strategy was developed for the discovery of prostate cancer markers in serum utilizing both animal models and human cohorts (Cima et al, 2011). Even though there are thousands of biomarker candidates discovered for cancer only (Polanski & Anderson, 2007), only few have been able to undergo the necessary validation steps for U S Food and Drug Administration (FDA) approval (Füzéry et al, 2013).

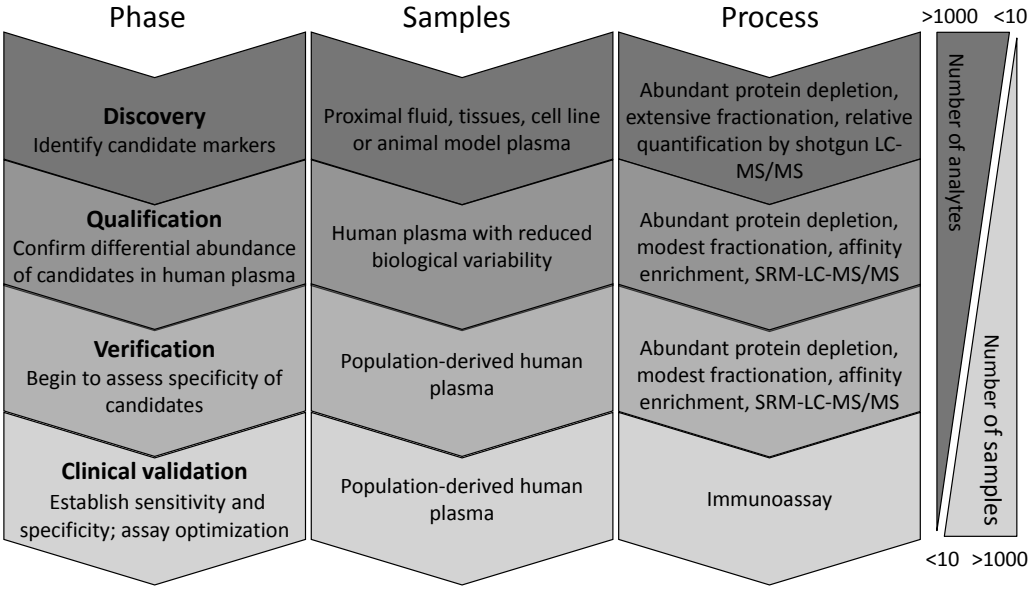

Figure 3. Biomarker discovery pipeline using MS (Rifai et al, 2006).

Discovery

A discovery phase usually consists of a data dependent shotgun proteomics experiment that is designed to identify and quantify as many proteins as possible in a small set of samples, as comprehensively and quickly as possible (Figure

3). In biomarker discovery these non-targeted, usually relative quantification techniques, report fold changes of proteins. Typically the MS-based relative quantification techniques begin with the digestion of the protein mixture and the actual quantification is subsequently performed based on the abundance of the unique peptides of this mixture. Important aspects to be considered in a discovery phase proteomics experiment are: 1) the number of samples, 2) sample collection and handling, 3) inclusion/exclusion criteria of the samples and 4) the possible downstream effects of sample preparation.

Discovery phase typically produce hypothetical biomarker candidates that need to be verified and validated, as with the current technologies low abundant proteins cannot always be reliably quantified (Rifai et al, 2006). The instrumental limitations affect the analysis and it has been also estimated that only approximately half or the cellular proteome is detectable with current MS-setups, while the other half of low abundant proteins raises challenges (Chick et al, 2015). Often the modified peptides from abundant proteins mask the unmodified peptides from low abundant proteins already at the ionization stage (Nielsen et al, 2006). Moreover, in clinical samples, proteins with specific disease-related mutations and proteins with PTMs or alternative splicing patterns are not efficiently monitored with these techniques, because of the inefficiency of database search algorithms to correctly identify them (Chick et al, 2015). However, it has been recently shown that by applying a mass-tolerant database search (Chick et al, 2015) or by using multiple search engines in the identification step (Shteynberg et al, 2013) the number of confidently identified PSMs can be increased substantially. Additionally, when precursor peptides are chosen for fragmentation in DDA, only the most abundant species are fragmented (Figure 4, Liu et al, 2004). Even though the repeated injections of the samples increase the number of new identified peptides, the instrument often fragments the same precursors with every subsequent injection (Tabb et al, 2010). Dynamic exclusion of recently fragmented precursors (Davis et al, 2001), scheduled inclusion lists for low abundant, unidentified precursors (Schmidt et al, 2008) and semi-targeted analysis of proteotypic peptides by accurate inclusion mass screening method (Jaffe et al, 2008) among other approaches (Bailey et al, 2014; Bateman et al, 2014) have successfully been used to improve this coverage (Figure 4).

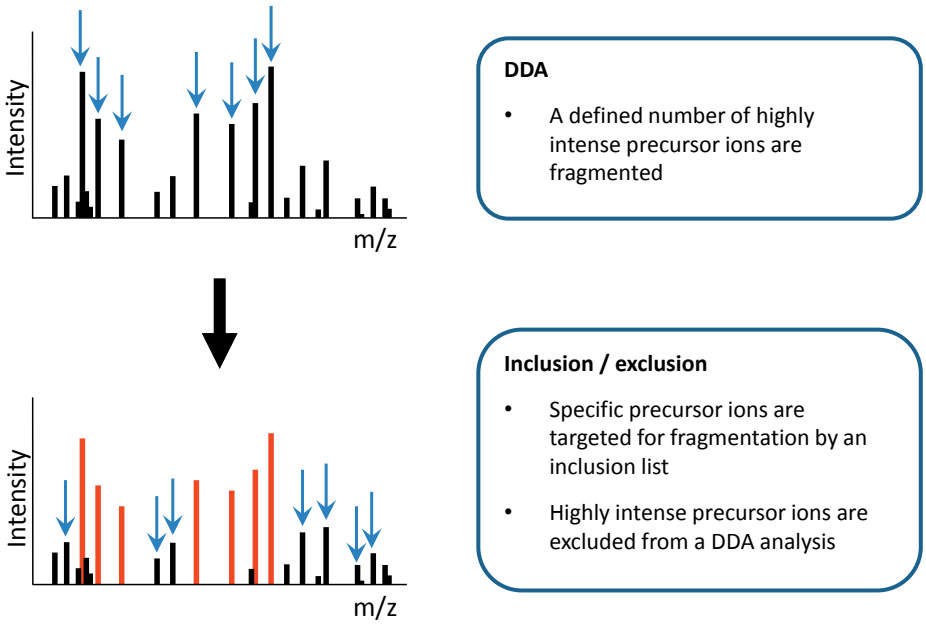

Figure 4. DDA, inclusion and exclusion approaches in increasing the proteome coverage in a complex sample. The low abundant precursor ions that are overlooked in DDA can be targeted either with a specific inclusion of low intensity precursors or by the exclusion of high intensity precursor ions. DDA, data dependent analysis; m/z, mass-to-charge ratio.

The efforts to discover biomarkers directly in plasma are hampered by a high dynamic range of different protein concentrations (Anderson & Anderson, 2002). Even though plasma is easily accessible and has the potential to reflect a disease state (Farrah et al, 2011), in the discovery phase of the MS based biomarker workflow, usually tissues or proximal fluids are analyzed. Plasma is analyzed later in the verification and validation experiments due to its complexity and the presence of highly abundant proteins. The current mass spectrometers in the DDA shotgun mode are capable of detecting peptides in four orders of magnitude (Gillet et al, 2012), whereas the plasma proteome consists of protein concentrations of 10 orders of magnitude (Anderson & Anderson, 2002), which is further increased by the digestion to peptides (Vandermarliere et al, 2013). Low level proteins in plasma can be enriched by extensive fractionation (Selvaraju & Rassi, 2012), specifically by the removal of high abundant proteins by immunodepletion (Zolotarjova et al, 2008), by other affinity based methods (Medvedev et al, 2012), by combinatorial peptide ligand libraries (CPLLs, Boschetti & Righetti, 2009) or by selective trypsin cleavage (Fonslow et al, 2013).

Currently, attempts to increase the coverage of low abundant proteins in a sample have been made. For example, using repeated depletion strategies and multiple

columns in concert in immunoaffinity, increased amounts of low abundant proteins were detected (Cao et al, 2013; Corrigan et al, 2011; Shi et al, 2012; Shuford et al, 2010). However, it has been noticed that the use of the tandem approach decreases the quantitative reliability of high abundant proteins (Patel et al, 2012). Additionally, others have observed that the sequential dithiotreitol-acetonitrile precipitation of high abundant disulfide-bond-rich proteins increased the identification rate of moderately abundant proteins (Fernández-Costa et al, 2012). However, although physicochemical depletion methods offer excellent reproducibility, it has been noticed that the most efficient removal of highly abundant proteins in plasma is achieved by immunoaffinity methods (Liu et al, 2011; Mahn & Ismail, 2011). Finally, even though the affinity based methods can reduce the dynamic range of protein concentrations and give possibility to detect low-abundance plasma proteins (Tu et al, 2010), also non-targeted proteins bound to abundant proteins are removed (Bellei et al, 2011; Patel et al, 2012). Indeed, it has been suggested that the proteins and peptides bound to albumin, for example, could be also a source of valuable diagnostic information (Petricoin et al, 2006). In contrary to the depletion methods, CPLLs are used to enrich the low abundant species. While providing comparable results to immunoaffinity depletion methods (Malaud et al, 2012; Milan et al, 2012; Millioni et al, 2011; Tu et al, 2011), CPLLs require relatively high amount of starting material, which restrict their use with clinical samples. It has been shown lately, however, that the process can be successfully down-scaled (von Toerne et al, 2013).

Complex datasets can cause biases and high FDRs in multiple levels starting from the choice of control and case groups, through sample preparation, analytical problems, database searches and statistics (Percy et al, 2013b). The preanalytical variability needs to be controlled throughout sample collection, storage and preparation and includes also variables considered in the study design such as lifestyle, medication, age and gender (Ahmed, 2009; Ferguson et al, 2007). The variability stemming from separate steps of typical proteomics sample preparation workflow can be substantial and represent up to 70% of total variance (Piehowski et al, 2013). The use of automation, standardizing procedures and reduction of the number of sample preparation steps has been suggested in order to reduce this variability (Krüger et al, 2013).

Analytical variability includes the variability stemming from the instrumental platform. This issue has been also been noted in various interlaboratory studies (Paulovich et al, 2010; Rudnick et al, 2010). All the sources of variability have an effect on the statistical

power of the analysis, which eventually determines the number of samples to be analyzed to reach statistical significance. A high number of different technical variables can be monitored during a LC-MS/MS run and multiple automated platforms have been released for this purpose. Depending on the platform, up to 284 different parameters can be monitored (Ma et al, 2012; Rudnick et al, 2010; Scheltema & Mann, 2012; Taylor et al, 2013). Examples of these metrics are: 1) The charge state distributions of ions, 2) ion injection times, 3) dynamic sampling, 4) MS1 and MS/MS intensities and 5) the total number of PSMs, peptides and protein groups. However, not all metrics are important for all assays or instruments. Also the choice of the standard depends on the application. Single peptides, peptide mixtures (Köcher et al, 2011a), protein digests (Bereman et al, 2014) complex tissue or cell lysates (Paulovich et al, 2010) or the sample itself (Rudnick et al, 2010) can be used as standards. As a rule of thumb, the more simple standards are generally good for extracting information on the LC-MS performance, whereas complex standards can generate knowledge on the whole workflow from sample preparation to data interpretation.

Qualification and verification

A good biomarker candidate needs to be evaluated for use in a clinical setting. The method designed for discovery needs to be optimized to verification (Figure 3). Prior to verification, however, candidates need qualification (Figure 3), as the discovery phase often produces too many biomarker candidates. Also, the analysis can be restricted to secreted proteins having a signal sequence, N-glycosylation site or a literature-based prediction to be secreted, therefore having a better probability to be found in plasma (States et al, 2006; Zhang et al, 2007). Those proteins showing the most significant difference in abundance between cases and controls can also represent general markers of stress or inflammation, for example, thus not specific and suitable diagnostic markers (Diamandis, 2004). Especially in these cases, the prioritization can be carried out based on previous biological knowledge on biological process being activated or certain class of proteins being tested (Goh et al, 2012). A hypothesis can also be made based on other omics datasets such as mRNA microarray or metabolomics.

The verification step (Figure 3) assures that the differential expression of the protein remains measurable in a new sample matrix using a novel assay and allows further refinement of the candidate list. The assessment of the specificity and sensitivity of each marker starts at this stage. These two measures define the diagnostic accuracy of the test: A sensitive clinical test can correctly identify patients with the disease

and a specific test can correctly distinguish those patients without the disease of interest (Abruzzo et al, 2015). The aim of verification is to determine the potential of a biomarker candidate to be used in the clinics, and thus, to be investigated further in validation phase. The approach used should allow the analysis of maximum number of analytes with high throughput at an affordable cost and produce small number of candidates meeting the standards of clinical validation (Anderson, 2005). Typically in the verification phase hundreds of individual patient samples will be used to assess the usefulness of tens of marker candidates.

A common problem concerning the large scale analysis of proteomics data is called overfitting; in overfitted data the results cannot be reproduced with a different set of samples. This problem is inherent to omics datasets, because in these technologies the number of data points always far exceeds the number of samples. To avoid overfitting, validation studies should be carried out in the verification phase of the MS-based biomarker discovery pathway (Boulesteix & Sauerbrei, 2011). In internal validation the study population is divided into two equal groups: a “training set” that is used to build a statistical model and a “validation model” that is analyzed parallel with the training set. In external validation a completely independent population is used to verify the results obtained in the discovery phase (Duffy et al, 2015). The training and validation sets should be independent and large enough to obtain reliable results (Ransohoff, 2004). A famous example of overfitting is a study conducted by Petricoin and co-workers (2002), where they claimed that with the SELDI method they could diagnose ovarian cancer from patient serum samples with 100% sensitivity and 95% specificity with a positive predictive value of 94%. However, these results could not be repeated by others and seem to hold true only in the dataset in question (Baggerly et al, 2004). Even though some of the problems in the example were related to the SELDI technology (Diamandis, 2003; Diamandis, 2004), it addresses the critical importance of independent validation in biomarker discovery.

The methods used in verification typically offer better protein quantification than in the discovery phase and the samples should represent the population the final clinical test is intended to be applied for. The verification phase has traditionally been conducted using ELISAs. However the lack of good quality antibodies for all possible candidates and the expense of generating novel antibodies have limited their use (Del Campo et al, 2015). Therefore, an SRM assay in a proximal fluid or in plasma is currently the favored

method in proteomics laboratories for this work (See chapter 2.2.3). The SRM approach offers the highest MS throughput, sensitivity, accuracy and possibility for multiplexing. The SRM method targets the analyte specifically and, depending on the labels used, provides absolute or relative quantification of the analyte concentration (Villanueva et al, 2014).

With extensive fractionation and sample pre-processing the sensitivity of SRM can reach 1-10 ng/ml (Hüttenhain et al, 2009), whereas the biologically relevant concentrations of plasma proteins can range from 50×10^6 ng/ml (serum albumin) to 5×10^{-3} ng/ml (interleukin 6; Anderson & Anderson, 2002). When stable isotope based SRM is combined with immunoaffinity purification by specific antipeptide antibodies (SISCAPA; Anderson et al, 2004), the sensitivity of the system can be further improved. Moreover, undepleted plasma has been increasingly analyzed, for example, a SRM assay of 142 plasma proteins ranging from 30×10^6 ng/ml to 44 ng/ml was reported (Percy et al, 2014). Zhang et al (2011) demonstrated that minimally labeled and label-free versions of SRM are cost-effective alternatives to full labeled SRM peptide pattern. A distinct advantage of SRM is its capability of multiplexing: With SRM, up to 100 candidate protein markers can be simultaneously analyzed in more than hundred patient samples. For example, 757 peptides were analyzed using 6050 transitions in a complex yeast lysate in one hour (Kiyonami et al, 2011). Additionally, samples from four different patients have also been analyzed simultaneously, by the use of iTRAQ reagents, during one SRM run (Yin et al, 2013)2013. Furthermore, the consistency of SRM results of seminal plasma, urine and serum have been validated in multi-laboratory studies and kits for SRM standardization have been evaluated on various platforms (Percy et al, 2013a; Prakash et al, 2012). Also SWATH has been used in biomarker verification studies, however, it offers 2-3 times lower sensitivity than SRM (Liu et al, 2013).

Clinical validation

At the biomarker validation phase (Figure 3) in a clinical setting, thousand or more patient samples are measured with CV values preferably lower than 10%. The validation phase requires high throughput together with accuracy and precision not available for MS based methods. This part of the work is typically performed using antibody based techniques, such as ELISA, even though the development of a novel antibody based assay is time consuming and expensive and cross reactivity is possible (Hoofnagle & Wener, 2009; Krastins et al, 2013; Wang et al, 2009). When establishing a clinically validated

test, biological variation (between and within subject variation), imprecision and bias should be calculated and remain within an accepted range (Fraser & Petersen, 1999). The linearity of the detection method and the limits of detection and quantification should be defined to outline the range of adequate precision and accuracy (Addona et al, 2009). Also reference intervals need to be established to distinguish healthy individuals from those affected (Solberg, 1987; Solberg & PetitClerc, 1988). Importantly, the issue of preanalytical variability needs to be controlled. This includes sampling, for example if analyte needs to be collected on a fasting state or stored in a special conditions, physiological factors such as age, gender or ethnicity and lifestyle factors, obesity and smoking.

At the clinical validation phase the clinical utility of a marker is established. Experience with current markers, that did not always undergo clinical validation, has highlighted the importance of this phase (Duffy et al, 2015; Stenman, 2013). One suggestion for this evaluation involves phase 1-4 trials, that are performed on independent study populations and highlight diverse performance features of the novel marker (Sackett & Haynes, 2002). The phase 1 (exploratory phase) determines whether the test can distinguish patients with confirmed condition to those not having the condition. If the area under curve (AUC) by ROC (Receiver operating characteristic) is below 0.7, the test is considered inapplicable for the disease (Obuchowski et al, 2004). The phase 2 (challenge phase) examines the accuracy of the test in a well-defined study population: Whether the presence or absence of the condition could be predicted by using distinctive thresholds for sensitivity and specificity. Subsequently, Phase 3 (advanced clinical phase) is used to establish the diagnostic accuracy and predictive values in the target population of the test (Obuchowski et al, 2004; Sackett & Haynes, 2002). Finally, Phase 4 (outcome phase) determines whether the test has an effect on the health of the patients tested. This involves the follow-up of the tested patients as well as those who did not undergo testing (Sackett & Haynes, 2002).

2.3.2 Examples of proteomes

Clinical proteome of endometriosis

Endometriosis is a condition where tissue resembling the uterine inner layer, endometrium, is found in pelvic cavity and/or in ovaries or on the surface of bladder or rectum. The prevalence of the disease in women in reproductive age is 2-10%

(Giudice, 2010) and up to 50% in women with unsolved infertility and/or pain symptoms (Eskenazi & Warner, 1997). Women with endometriosis suffer from chronic pelvic pain caused by innervation and inflammation at locations of endometriotic lesions (Berkley et al, 2005; Tokushige et al, 2006). The etiology of endometriosis is not well understood and currently both the clinical diagnosis and the treatment of endometriosis involve surgery. Among the hormonal treatments of endometriosis are e.g. the contraceptive pills and estrogen blocking gonadotropin-releasing hormone (GnRH) analogs. However, these treatments are not always effective or bring up severe side effects (Giudice, 2010). Therefore, new information on possible therapeutic targets and diagnosis for endometriosis would be essential. Moreover, the progression of endometriosis is presently impossible to predict. A biomarker for endometriosis would be especially useful in cases, where the disease is progressing fast. Also, as the etiology of endometriosis related infertility is currently poorly understood, and treatments either aim at the alleviation of the pain symptoms or the treatment of infertility, a marker estimating the likelihood of infertility would be very welcome for many endometriosis patients.

The complexity of different types of endometriosis as well as the heterogeneous nature of the endometrium tissue along its modulation with the menstrual cycle makes the disease difficult to study (Bulun, 2009; Farquhar, 2007). Many endometriosis studies have aimed at understanding the biology leading to the pathophysiology of the disease, but also in hope to find new targets to medical treatments and diagnostics. The proteome of endometriotic tissue has been studied during the years using various patient material, sample types and methods (Table 2). However, these methods are partially outdated and do not offer a direct molecular description of the disease. Indeed, in the studies patients are often compared to a control population relying on visual comparison of 2-DE gels, DIGE or SELDI providing patient-control distinction based on unidentified peptide peaks (Table 2; May et al, 2011). Only very recently, an attempt of more comprehensive quantitative definition of endometriosis proteome has been made (Table 2; Kasvandik et al, 2015). Nevertheless, most studies have been performed in small-scale, studying one molecule or a small number of molecules. Extraction of conclusive data from these results remains difficult, as in many of the studies most of the peaks reported in endometriosis samples were not subjected to MS-based identification. Therefore, no promising biomarkers of endometriosis were identified in these studies.

Table 2. Published studies on the proteome of endometriosis. Drawn according to respective publications.

Sample	Method	Coverage	Reference
Endometrium tissue	2-DE	11 spots	(Zhang et al, 2006b)
Serum	2-DE	13 spots	(Zhang et al, 2006b)
Endometriotic tissue	SELDI	41 peaks	(Kyama et al, 2006)
Endometrium tissue	2-DE	~100 spots	(Fowler et al, 2007)
Endometrium tissue	2-DE	119 spots	(Ten Have et al, 2007)
Serum	SELDI	20 peaks	(Liu et al, 2007)
Endometrial fluid	2-DE	52 spots	(Ametzazurra et al, 2009)
Serum	SELDI	57 peaks	(Seeber et al, 2010)
Endometrium tissue	DIGE	36 spots	(Stephens et al, 2010)
Endometrium tissue	SELDI	36 peaks	(Fassbender et al, 2010)
Endometrium tissue	2-DE	70 spots	(Rai et al, 2010)
Urine	Weak cation exchange, MALDI-TOF	6 peaks	(El-Kasti et al, 2011)
Urine	2-DE	133 spots	(Tokushige et al, 2011)
Serum	DIGE	25 spots	(Faserl et al, 2011)
Serum	Weak cation exchange, MALDI-TOF	46 peaks	(Zheng et al, 2011)
Peritoneal fluid	2-DE	11 spots	(Wölfler et al, 2011)
Endometrium tissue	SELDI	5 peaks	(Fassbender et al, 2012a)
Serum	SELDI	92 peaks	(Fassbender et al, 2012b)
Endometriotic tissue	2-DE	50 spots	(Hwang et al, 2013)
Serum	Weak cation exchange, MALDI-TOF	13 peaks	(Long et al, 2013)
Follicular fluid	2-DE	29 spots	(Lo Turco et al, 2013)
Endometriotic tissue	iTRAQ	419 proteins	(Marianowski et al, 2013)
Urine	Weak cation exchange, MALDI-TOF	36 peaks	(Wang et al, 2014)
Serum	2-DE	7 spots	(Hwang et al, 2014)
Serum	2-DE	13 proteins	(Tuerxun et al, 2014)
Serum	DIGE	95 spots	(Dutta et al, 2015)
Omental fat	iTRAQ	3 proteins	(Williams et al, 2015)
Endometriotic tissue	Weak cation exchange, MALDI-TOF	8 peaks	(Zhao et al, 2015)
Endometrial and endometriotic cells	Super SILAC	6900 proteins	(Kasvandik et al, 2015)

The studies with endometriosis are affected by preanalytical variability that is related to the complex disease itself. First of all, the classification system of the American Society of Reproductive Medicine (ASRM, 1997) often used for estimating the severity of endometriosis does not work efficiently in estimating the pain symptoms of patients (Haas et al, 2013; Vercellini et al, 2007). Thus, clinically significant disease is difficult to correlate with symptoms (Eskenazi & Warner, 1997; Waller et al, 1993). Therefore,

a patient classification system based on lesion location and appearance has been suggested, even though one patient has often more than single type of lesion in multiple sites (Adamson et al, 1982; Koninckx & Martin, 1992). In addition, the symptoms of a patient, ethnicity and demographic and reproductive factors of endometriosis patients need to be considered as conflicting results from multiple populations has been received in prior studies (Rogers et al, 2013). Also the use of hormonal contraception has been shown to affect serum biomarker profiles (de Sá Rosa e Silva et al, 2006; Piltonen et al, 2012). Importantly, the phase of the menstrual cycle, and features such as cycle length and regularity should be controlled as these might have an effect of gene and protein expression (Fassbender et al, 2012a).

Variability in the collection, processing and storage of specimens needs to be controlled as this can cause a considerable bias or errors in measurements. Sample handling alone can, thus, prevent an identification of a disease related molecular changes (Ransohoff & Gourlay, 2010; Tworoger & Hankinson, 2006). For endometriosis studies, eutopic endometrium, myometrium, peritoneum and the ectopic endometriosis tissue are often collected. The ectopic endometrium is a heterogeneous tissue that contains stromal and glandular epithelial cells as well as possible inflammatory, fibromuscular and blood cells. Therefore, the acquired tissue specimen needs to be studied by a pathologist before further use. Analysis should include the estimation of the percentage of glands, stroma and sites of inflammation. The tissues shoul be kept cool at all times to avoid molecular degradation. Also the collection should rather be performed by sharp dissection than by electrosurgery and the excised tissues need to be stored as soon as possible (Sheldon et al, 2011). Even though recently formalin fixed paraffin embedded tissues have been successfully analyzed by proteomics methods, long storage of the tissues in -80 °C or liquid nitrogen is recommended (Becker, 2015; Ralton & Murray, 2011).

Plasma, serum, urine, saliva, peritoneal fluid and endometrial fluid aspirate are the most common biological fluid samples used in endometriosis research. Blood is the most used bio specimen because of relatively easy collection from both patients and control subjects. Blood and its derivatives, plasma and serum, consist of complex matrix of biomolecules that reflect various biological processes in the body. Therefore, the disease-related effect in these matrices may be diluted or disguised. Saliva is used to monitor DNA and hormones (Koni et al, 2011; Shirtcliff et al, 2001). It may, however, be contaminated with bacterial proteins (Abraham et al, 2012). Peritoneal fluid is of special

interest in endometriosis studies, as it is found in close proximity of endometriosis lesions (Koninckx et al, 1999). However, the collection of the fluid requires surgery, the complexity is similar to that of plasma and the fluid volume varies within the menstrual cycle (Pupo-Nogueira et al, 2007). Furthermore, it has been recommended that the peritoneal fluid sample would be collected before the excision of the ectopic endometriosis tissues to minimize the contamination of the sample with blood, cyst fluid or tissue.

Liver proteome of aromatase overexpressing mouse model (AROM+)

Aromatase overexpressing mouse model (AROM+) was generated to study the significance of estrogen-androgen ratio in male reproductive functions (Li et al, 2001; Li et al, 2006; Li et al, 2004; Li et al, 2002). The human aromatase enzyme is responsible of converting androgens to estrogens. AROM+ mice have a ubiquitously expressed human P450 aromatase transgene, which results in elevated serum estradiol and prolactin concentrations, as well as decreased levels of testosterone in male mice. Whereas the phenotype of female mice is unchanged, male mice with aromatase overexpression have multiple defects in the reproductive organs, causing infertility and other consequences (Li et al, 2001; Li et al, 2006; Li et al, 2004; Li et al, 2002). The function, and thus the proteome of AROM+ liver is of specific interest as liver steatosis together with decreased fatty acid oxidation and increased plasma lipids has been reported in aromatase knockout mice (ArKO; Nemoto et al, 2000).

Liver is a multifunctional organ the main roles of which are to process digested nutrients, detoxify substances and to produce plasma proteins. The main obstacles in liver proteomics studies have been the large quantity of abundant metabolic enzymes that can mask lower abundant proteins in the analysis. The liver also has an active circulation of blood that complicates the quantification of proteins expressed in liver. However, liver is relatively homogeneous as tissue and has been extensively studied using different proteomics methods (Low et al, 2013; Shi et al, 2007; Sun et al, 2010; Yan et al, 2004). A reasonable high availability of the sample makes mouse liver tissue an attractive target for in-depth protein identification and quantification experiments.

Mouse liver has been studied with LC-MS label-free quantification methods. Cutillas and Vanhaesebroeck (2007) analyzed multiple mouse tissues in order to create a quantitative protein catalogue for mouse brain, heart, kidney, liver and lung. In this

work the samples were first separated by SDS-PAGE, then analyzed by LC-MS/MS on 11 fractions and subjected to label-free quantification (Cutillas & Vanhaesebroeck, 2007). In another study, a label-free quantitative platform was validated by liver hepatocyte analysis of a liver-specific adenomatous polyposis coli (APC) knockout mouse. Also in this study, the sample was divided into fractions by SDS-PAGE analysis and 1789 quantifiable proteins were reported (Vasilj et al, 2012). More recently, after a hepatocyte compartment specific fractionation in BIRC5-knockout mouse model, between 2289 and 1717 proteins, depending on a fraction, were identified and could be used for label-free quantification (Bracht et al, 2014). The group of Mann recently published an in-depth profile of mouse liver and several hepatic cell types. After extensive fractionation and hundreds of hours of LC-MS/MS, over 11 000 proteins were identified and quantified comprising the largest tissue proteome database to date (Azimifar et al, 2014). After its publication this resource has been used to map conserved small open reading frames in mouse genome (Mackowiak et al, 2015), to estimate the copy number distribution of different transcription factors in eukaryotic transcription (Rybakova et al, 2015) and to study stress reactions in murine hepatocytes (Marcelo et al, 2015).

2.3.3 Integrating proteomics with other large-scale data

There are sophisticated methods to generate data at each molecular level, which can be combined to understand a disease or other phenotype alterations. Technological developments in DNA sequencing, mRNA based work and epigenetics (Buermans & den Dunnen, 2014), proteomics (Breker & Schuldiner, 2014) and metabolomics (Fuhrer & Zamboni, 2015) have created a demand to analyze these large amounts of data coordinately to produce complementary results describing the biological system from a molecular perspective. However, as omics technologies are notorious for generating false positive readings, the combination of different data from different sources can increase the reliability of the results, in addition to use of validation sets and proper statistical methods already during discovery stage. If multiple sources of information indicate the same activated pathway, it is less likely that the result is a false positive finding. There are multiple levels of deregulation in the system, which is missed when analyzed by a single omics method only. Moreover, the possibility of modelling an entire system from genomic DNA to mRNA (including splicing) to histone modifications to proteins to protein modifications and to metabolites would

expand the knowledge of the molecular mechanisms and causal relationships among the molecules.

Systems biology has taught us that integration is needed, as it is clear that not only genetic mechanisms affect the phenotype, in a specific disease for example. Multiple studies have confirmed that the correlation between protein abundance and transcribed mRNA is often modest (de Sousa Abreu et al, 2009; Nesvizhskii et al, 2006; Ning et al, 2012; Schwanhäusser et al, 2011; Schwanhäusser et al, 2013; Vogel & Marcotte, 2012), which highlights the transcript-independent protein regulation in a dynamic biological system. Commonly protein abundance and mRNA expression are used to complement each other to avoid technological bias and to achieve a better understanding of the global events in a cell or a tissue. Comparative studies of transcriptomics and proteomics offer an opportunity to study the quantitative transcript-translated protein relationships, protein regulation and translational regulators. However, also new tools for assessing the interactions between the transcriptome, proteome and metabolome are needed.

The approaches in the integration of transcriptomics, proteomics and metabolomics data, for example, can be roughly divided into two: 1) Multi-staged analysis that is hierarchical and performed stepwise and 2) meta-dimensional analysis, where the data is first integrated and subsequently formed into a multivariate model. In the multi-staged analysis mRNAs and proteins are considered to serve adjacent functions and protein profiles reflect mRNA profiles. The parallel comparison of the two should give a more comprehensive picture of protein and gene expression, as proteomics or transcriptomics alone cannot sufficiently describe a biological system. Multi staged analysis relies on an assumption that a phenotype stems from a linear association between genetic information and proteins, for example (Figure 5). When this approach is used, the data is first subjected to initial analyses that are then hierarchically integrated together with the reduction of data at each step. Functional and pathway information can be used as an additional step in the multi-staged analysis, where only identifiers, in one or multiple omics datasets that have certain functional annotations are studied further (Holzinger & Ritchie, 2012).

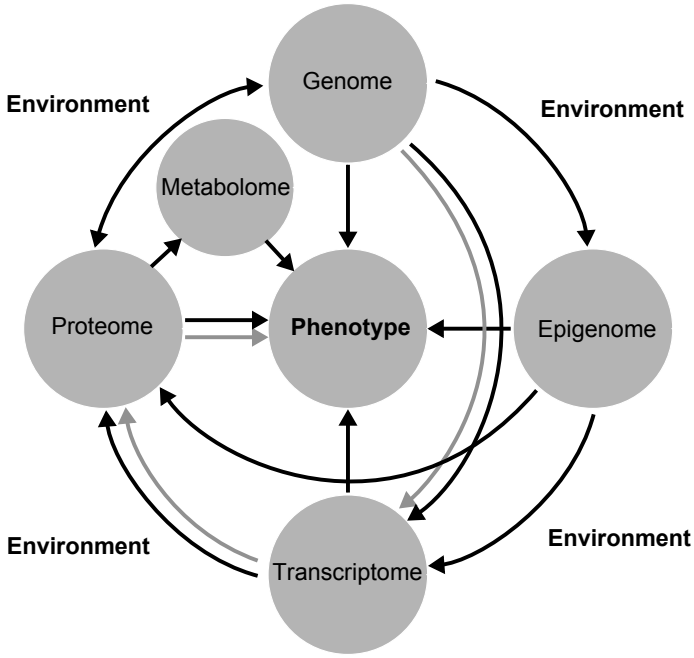

Figure 5. Different approaches to data integration. The multi staged analysis relies to the assumption that the phenotype is a result of a linear association between genome, transcriptome and proteome (gray arrows), whereas the meta-dimensional analysis is based on the concept of interconnected network of associations (black arrows). Modified from Richie and coworkers (2015).

The expression of a gene often does not correlate with the amount, or predict the location or activity of a protein. This is due to the post-transcriptional, translational and post-translational regulation within or outside the cell. Also biochemical control occurs, such as allosteric or feedback regulation of enzymes. On the other hand, especially post-translational modifications and allosteric regulations affect the quantities of different proteins within a cell and their relations to gene expression and metabolome (Güell et al, 2011; Weiss & Atkins, 2011). Thus, in the meta-dimensional analysis, the ground assumption is that different layers of molecular information, such as gene transcripts, proteins and metabolites create an interconnected network without strictly linear relationships (Figure 5). This hypothesis requires data integration prior to further analysis, which generates multiple challenges: First, the different omics datasets have variable sizes of databases, the patterns and types of missing data vary, and there are different levels of noise, correspondence and correlation. Multiple approaches to combine the data have been developed, such as concatenation, transformation and model-based integration schemes (Holzinger & Ritchie, 2012; Kim et al, 2012).

The integration of omics data encompasses multiple data related challenges. The omics data files are typically large and require computational power to be integrated and analyzed (Berger et al, 2013). The quality is usually controlled by the use of summary statistics; due to the very high number of variables the quality assessment of each data point is not possible. The data overviews are performed often at sample level, by the comparison of different distributions or extracting single variables suited for quality control (Zuvich et al, 2011). The scale and dimensionality of the data generates statistical challenges. The high number of variables together with a reasonable low number of samples generates limited statistical power as the number of tested hypotheses is vast. The use of a data reduction strategy is, therefore, often necessary to create a refined subset of data (Johnstone & Titterington, 2009). Approaches to data refinement can be extrinsic, such as literature based filtering, or intrinsic, where an unsupervised method is used to find subgroups within the data itself, such as principal component analysis (PCA) (Bonner et al, 2014). The integrated data can also be very heterogeneous and have confounders that have an effect on the statistical analysis. The attempts to overcome the issue include the exclusion of confounders and a surrogate variable analysis (Leek & Storey, 2007). The experiments are also often replicated to check for false positive identifications and functional validation by the use of an orthogonal method (Van Poucke et al, 2012), text mining (Raychaudhuri et al, 2009) or in silico models (Crooke et al, 2006) has also been recommended.

Concrete examples of studies successfully combining proteome data with other large-scale datasets have been published lately. Wang et al (2011b) identified significantly elevated complement proteins in the synovial fluid of osteoarthritis patients. Based on subsequent genomics and immunohistochemical analyses, the source of proteins were confirmed to be from the synovial membrane. Furthermore, the key proteins directly involved in the pathogenesis of osteoarthritis were identified using murine knock-down models (Wang et al, 2011b). Another group used information from genome-wide association studies that predict that certain mutant proteins could be the key molecules in the process of cancer transformation and progression in different tumors (Wang et al, 2011a). A SRM assay was developed for these proteins, which was tested in both cancer cells and clinical samples of colorectal and pancreatic cancer. The method was found to be suitable of detecting the relative abundance of these proteins in cancer cell lines and in clinical samples (Wang et al, 2011a).

3. AIMS OF THE STUDY

When correctly applied and optimized for a sample matrix, modern shotgun proteomics-based label-free mass spectrometry (LC-MS) methods are capable of identifying and quantifying thousands of proteins in complex samples. In this study, these methods were optimized in terms of sensitivity and robustness and then applied for aromatase overexpressing mouse liver and ovarian endometriosis tissue samples to study the changes in tissue proteome. Statistical methods were refined for the study data, and microarray and metabolite results were integrated in this information.

One of the goals of this research was to develop and demonstrate the ability and capacity of LC-MS for quantification of complex samples. With these tools in hand an overall aim was to establish and optimize workflows to extract biologically meaningful systems-level information from a mouse model and from a complex human disease.

Specific objectives for the study were:

1. To discover the best statistical LC-MS method for label-free quantification of proteins in a complex sample.

2. To convert quantitative proteome data from aromatase overexpressing mouse liver to meaningful biologically relevant information together with microarray and metabolomics information.

3. To study endometriosis disease using quantitative LC-MS of proteins to find ovarian endometrioma specific proteins that can be used as diagnostic markers.

4. MATERIALS AND METHODS

4.1 Methods used in this study

Method	In original publication
Direct tissue digestion	III
Protein precipitation and in-solution digestion	I, II
LC-MS/MS	I, II, III
AIMS	III
Inclusion lists	I, II, III
Sequence database searches	I, II, III
LC-MS label-free quantification	I, II, III
Pathway analysis	II, III
mRNA microarray analysis	II, III
Targeted metabolomics analysis	II
qRT-PCR	II
Western blot	III
Immunohistochemistry	III
General bioenergetics screening	II
Steroid hormone assays	II

Selected methods are described in detail below.

4.1.1 Direct tissue digestion

Frozen tissue sections of 10 µm in thickness were cut on a cryostat (Leica Microsystems) at −18 °C. Adjacent sections were either used for proteomic experiments or stained with hematoxylin and eosin. The size of the stained sections was measured using a stereo microscope (Zeiss SteREO Lumar V12). The frozen section was subjected to trypsin digestion by adding 5 ng of trypsin/mm^2 in a solution of 12 ng/µl trypsin in 30% acetonitrile (ACN) in 100 mM NH_4HCO_3, spinning the tissue at the bottom of the tube and incubating overnight at 37 °C. The digestion was performed without denaturation, reduction or alkylation reactions. After the digestion, the samples were centrifuged at 14 000g for 45 min and the peptide containing supernatant was collected and aliquoted. The supernatant samples were then lyophilized using a vacuum centrifuge (Hetovac, Heto Holten), and stored at −70 °C until use.

4.1.2 LC-MS/MS

The peptide samples were subjected to liquid chromatography –tandem mass spectrometry (LC-MS/MS). The analyses were performed on an ESI-hybrid Ion Trap-Orbitrap mass spectrometer (LTQ Orbitrap Velos/Velos Pro; Thermo Fisher Scientific) coupled to Easy Nano LC liquid chromatography system (Thermo Fisher Scientific). The LC-MS/MS system was controlled by Xcalibur software (Thermo Fisher Scientific). The peptides were separated using a reverse phase chromatography columns. The peptides were first concentrated by a 2.5 cm-long trap column and then separated by a 15 cm-long analytical column with an inner diameter of 75 µm. The columns were packed with 5 µm particle and 200 Å pore size C18 resin. The elution was accomplished with a flow rate of 0.3 µl/min and a 75 to 110 minute gradient, depending on a sample, from buffer A (98% H_2O, 2% ACN, and 0.2% HCOOH) to buffer B (95% ACN, 5% H_2O, and 0.2% HCOOH) by gradually increasing the content of organic solvent in the mobile phase. The MS/MS was performed in DDA mode using CID with TOP 15 or TOP 20 mode. The Orbitrap was operated at the mass range of 300-2000 m/z and a resolution of 60 000.

4.1.3 Sequence database searches

The database searches were accomplished using both Mascot and Sequest search algorithms available in the Proteome Discoverer search platform (Thermo Fisher Scientific). The spectra were search against a UniProtKB/Swiss-Prot human, mouse or yeast database including common contaminating proteins from cRAP (the common Repository of Adventitious Proteins). The search was limited to tryptic peptides with maximum of two missed cleavage sites fulfilling the criteria of precursor mass tolerance of 5 ppm and fragment mass tolerance of 0.5 Da. In those samples that were reduced and alkylated, methionine oxidation was selected as a dynamic and cysteine carbamidomethylation as a fixed modification. Accepted false discovery rate determined by Percolator decoy search was set to 1%.

4.1.4 Label-free quantification

All spectral data files were imported to Progenesis LC-MS software (Nonlinear Dynamics) for quantitative analysis. The LC-MS maps of the entire dataset were overlaid and aligned by placing a number of manual vectors followed by an automatic alignment. The peak picking and feature detection was performed in default sensitivity mode including singly, doubly and triply charged precursors with an elution time window of 12 seconds

and larger. The protein matches, comprising of proteins with unique peptides and two or more PSMs, were imported to Progenesis and integrated to the feature map. The peptide features were further refined by removing peptide contaminants. For a subset of data, a requirement of two or more identified spectra per feature with a precursor tolerance of 5 ppm was set. The normalization and statistical evaluation of the results was either performed in Progenesis or externally using other software packages, such as R.

5. RESULTS

5.1 Evaluation of LC-MS-label-free analysis (II, III)

During this thesis work, liver tissues from aromatase overexpressing (AROM+) male mice, and ovarian endometrioma (PO) and endometrium (PE) tissues from endometriosis patients were studied using quantitative LC-MS. Very similar MS and quantification workflows were used when the different tissue samples were analyzed, which offers a good opportunity for the evaluation of the technical performance of the analyses. Even though the analysis pipelines were similar, some sample related differences existed in the datasets. Therefore, to highlight the differences and similarities between the AROM+ and endometriosis datasets, the results associated to technical performance of the LC-MS quantification are described and evaluated in the following section.

Sample preparation

Sample preparation is a key step in any successful quantitative proteomics experiment. As label-free quantification methods are extremely sensitive for bias in sample handling and preparation, this work aimed at sample preparation with minimal sample handling steps, minimizing the chance for errors. For endometriosis tissues, a direct tissue digestion approach was chosen, where the trypsin-containing solution was deposited directly onto the tissue sections and the peptides were extracted by a centrifugation step after incubation. This simple sample handling protocol enabled us to extract 24-93 µg of peptides in endometriosis tissue samples ranging from 8.8 to 59.1 mm^2 in size, without homogenization, protein precipitation or peptide clean-up steps. On average 1.5 $\mu g/m^2$ (SD=0.6) of peptides was extracted from each tissue. In contrast, as larger amounts of tissue material was available, a more conventional sample preparation protocol was chosen for the AROM+ male mouse liver tissues. The sample preparation of these tissues included tissue homogenization and protein precipitation by acetone, as well as a salt removal step after trypsin digestion.

DDA analysis revealed over 10 000 peptide sequences

LC-MS/MS analysis is a powerful tool to identify thousands of proteins from a single tissue sample. In order to reach high proteome coverage in the liver and endometriosis

tissues, all samples were subjected into a standard shotgun analysis workflow consisting of a DDA of 15 most intensive precursor ions in each cycle. In the LTQ Orbitrap Velos, the duty cycle for this set up was 1.2 seconds, which allowed the acquisition a minimum of 10 MS/MS spectra over an eluting precursor peak. As a result, in the endometriosis dataset, on average over 3700 distinct doubly and triply charged peptide sequences were identified in each sample. For the liver dataset this number was significantly higher, being on average over 8500 identified peptide sequences in each sample. However, when the number of unique peptide sequences is calculated for the whole dataset, including 13 samples for endometriosis and 12 samples for AROM+, the difference is reduced: 10794 unique sequences were identified for all endometriosis and 12284 sequences for all of the liver samples.

Accurate inclusion of 18 candidate biomarkers in endometriosis

Even though the shotgun proteomics is effective in identifying a high number of peptide sequences in complex samples, many low abundant precursor ions are not subjected to fragmentation in DDA (Figure 4). Therefore, to increase the detection of low abundant peptides from biomarker candidate proteins in endometrium and ovarian endometrioma tissues, AIMS (accurate inclusion mass screening, Jaffe et al, 2008) was applied. In AIMS, a semi-targeted methodology is used, where the mass spectrometer is guided to analyze peptide sequences specific for a protein of interest, called proteotypic peptides. In the endometriosis study, an in-silico digestion of 18 proteins of interest to their corresponding peptide sequences was performed. This resulted in 1079 precursor ions of unique peptide sequences with charge states of 2+ and 3+. The instrument was then programmed to analyze only these masses in endometriosis tissue samples, resulting in 132 proteotypic peptide identifications representing 12 proteins of interest. However, the semi-targeted analysis approach based on m/z values was found to be reasonably inefficient, as 80 % of all identified sequences did not originate from the targeted proteotypic peptides.

Inclusion list approaches increased the number of identified peptide sequences up to 30 %

It has been estimated that a large part of the peptide signals detected by MS in the DDA mode remain unidentified due to the complexity of the sample (see section 2.3.1). Therefore, in addition to AIMS (see previous section), a more

general inclusion approach was applied to target these masses in both AROM+ and endometriosis datasets. In the directed mass spectrometric approach (Schmidt et al, 2008), those low abundant precursor ions that are not chosen for fragmentation during DDA, but have been detected at the MS-level, are chosen for re-analysis via inclusion lists (Figure 4). These inclusion lists are based on the RT and m/z of a peptide already detected, which increases the specificity of the identification in comparison to AIMS.

In terms of inclusion list construction, however, marginally differential procedures were chosen in the endometriosis and AROM+ datasets. In the endometriosis dataset the proteotypic peptide RT and m/z values from the AIMS experiment (see section 5.1.3) and all non-identified precursors (RT and m/z) from DDA (see section 5.1.2) were transferred to an inclusion list for the replicate injection a sample. In the AROM+ dataset all the non-identified precursors (RT and m/z) or precursors identified with only one PSM (see section 2.1.3) in the DDA experiment (see section 5.1.2) were reanalyzed in a mixture of peptides comprising of all the liver samples in the sample set. This mixed sample was then analyzed sequentially four times, always including only those precursors (RT and m/z) previously unidentified.

As a result, when each sample was studied independently, the use of inclusion lists increased the number of unique identified sequences (from doubly and triply charged peptide ions) on average by 8 % in both endometrium and endometrioma samples. Interestingly, this increase was lower in the liver sample (4 %, Figure 6). However, when all of the samples in endometriosis dataset were studied together, 12 % increase in identifications was recorded. When the inclusion of the proteotypic peptide sequences from the AIMS experiment (see section 5.1.3) was studied independently, 19 % increase in the identifications was found. Interestingly, in the complete AROM+ dataset, 30 % increase in identifications was detected. Accordingly, by the use of the DDA and inclusion list experiments, 2115 unique protein identifications were made in the endometriosis dataset and 1850 proteins were identified in the AROM+ liver dataset (Figure 7).

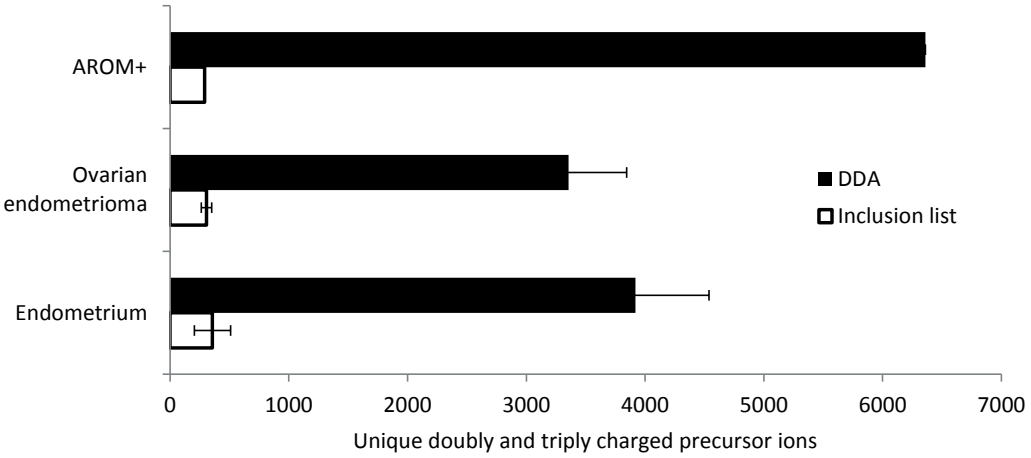

Figure 6. The number of unique 2+ and 3+ precursor ions identified in the mixture of all AROM+ samples, in ovarian endometrioma samples and in endometrium samples using DDA and directed mass spectrometry approach.

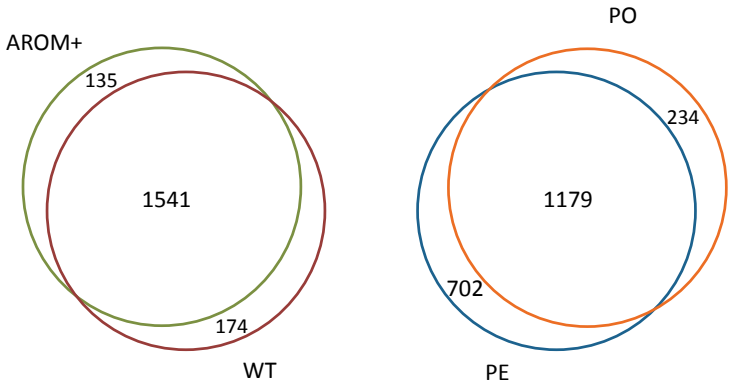

Figure 7. The number of proteins identified in AROM+ and in endometriosis datasets.

Quantification of proteins in endometriosis and AROM+

After the LC-MS/MS analysis, the peptide spectral data of both studies was imported into the Progenesis software for quantification. In Progenesis, all spectral files in the dataset are aligned into a single LC-MS map based on the m/z and RT values of the precursor ions. In this fashion the corresponding abundances of each individual precursor ion can be extracted over the whole dataset. After the alignment and feature detection (see section 2.2.2), Progenesis could define 7234 distinct peptide features associated with 1516 proteins in the entire endometriosis dataset and 11487 peptide features with 1499 linked proteins in AROM+ liver. As defined by repeated injections of a standard sample, in both cases the LC-MS system was operating with an analytical coefficient of

variance (CV) of 10% calculated after normalization for the quantitative data analysis (Figure 8). Moreover, when the sample sets were analyzed by a t-test with a threshold of at least three identified peptides per protein, q<0.01, power>0.8 and fold change >1.5, 214 differentially expressed proteins were found between patient endometrium and ovarian endometrioma and 93 proteins between WT and AROM+ liver (Figure 9). Interestingly, in the endometriosis dataset, all of the 12 biomarker candidates targeted in AIMS were found among the 214 differentially expressed proteins.

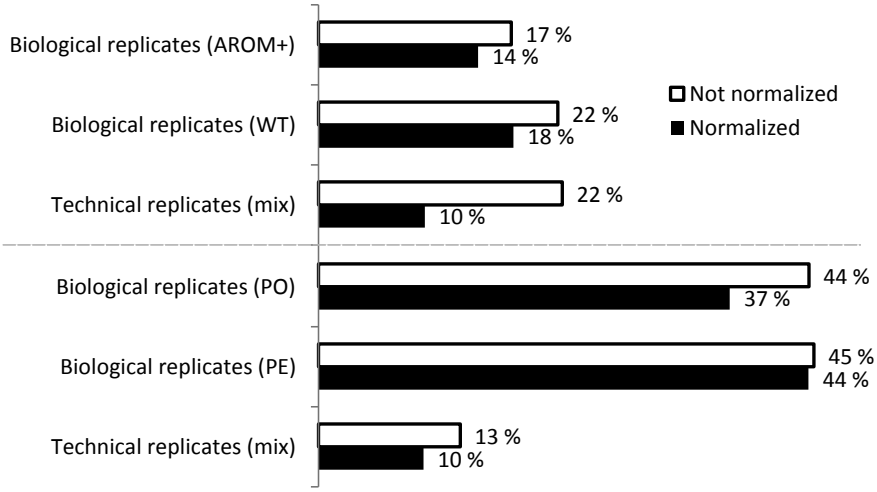

Figure 8. Median CV % of proteins identified in AROM+ and endometriosis datasets, before and after Progenesis-based normalization.

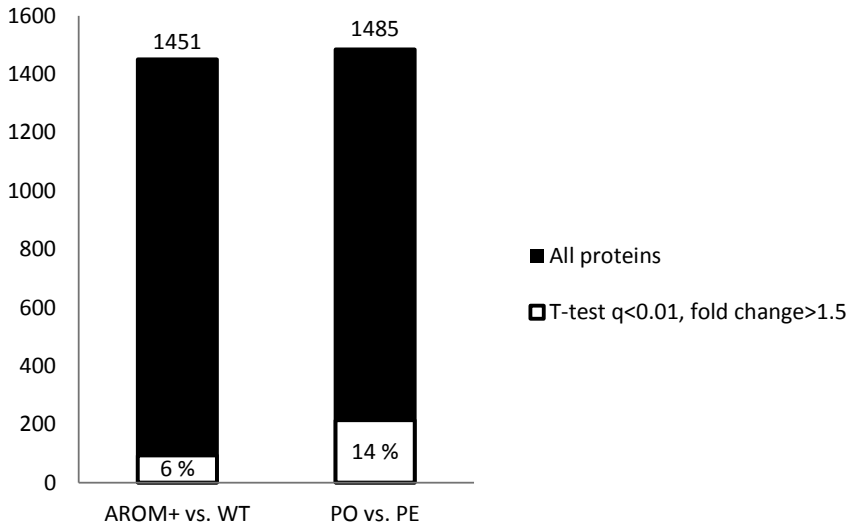

Figure 9. The percentage of differentially expressed proteins in AROM+ and endometriosis datasets.

5.2 The impact of statistical methods on quantitative proteomics data (I)

The performance of different statistical methods to label-free LC-MS quantification data was tested to find the best statistical approach for the quantification of differential protein levels in complex samples. The study was performed by adding a standard mixture of 48 human proteins (UPS1) at five different concentrations (2, 4, 10, 25 and 50 fmol/µl) to a complex yeast digest. The samples consisting of mixtures of human and yeast proteins were then subjected to shotgun proteomics analysis and LC-MS label-free quantification. As a result, after the LC-MS/MS analyses, 9953 unique peptide sequences were identified. These sequences originated from 1850 unique proteins including 47 proteins from UPS1 fulfilling the criteria for acceptable identification (see chapter 4.1.3). The UPS1 peptides were well represented in the dataset; on average 21 peptides were identified for each UPS protein, whereas generally five peptides per protein were identified in yeast. After Progenesis analysis, 944 proteins with at least two identified peptides were considered for further statistical evaluation, including those 46 UPS1 proteins identified with two peptides or more.

ROTS offers high sensitivity and specificity

The quantified UPS1-yeast data was subjected to statistical evaluation using five different approaches. The methods tested included a t-test, Significance Analysis of Microarrays (SAM), Linear Models for Microarray Data (LIMMA), Rank Product (RP) and Reproducibility-Optimized Test Statistic (ROTS). First, the sensitivity and specificity of the different methods were studied using the AUC values in ROC plots. Similar AUC values were received for the different concentrations of UPS1, except for the 2 vs. 4 fmol/µl comparison, which was presented with low sensitivity and specificity throughout the dataset. In respect to the different statistical methods tested, ROTS, RP and LIMMA provided high AUCs, t-test moderate values and SAM exhibited the poorest performance in the ROC analysis. To evaluate the proportion of differentially expressed UPS1 proteins in each evaluated method, a FDR level of 0.05 was used. Accordingly, the median sensitivity, i.e., the capability to detect the upregulated UPS1 proteins correctly, was high in most of the statistical methods tested, being 97.8% for ROTS, LIMMA and SAM. For the t-test and RP the median sensitivities were 91.3 % and 71.7 %, respectively. Of the methods, RP showed low sensitivity throughout, whereas the other methods performed poorly in the lower or higher range of spiked concentrations (2 fmol/µl and 50 fmol/µl) only.

Additionally, most of the methods demonstrated high number of false positives. The highest number of these background yeast proteins reported as changed (FDR<0.05) was

recorded with SAM and the lowest number with RP. It was concluded that the number of false positive detections could be reduced by the use of alternative normalization techniques, such as median scaling. Correspondingly, when a fold change threshold of two was used, the number of false positives was reduced without much reducing the number of true positive detections. Moreover, when the minimum number of peptides that were used for quantification was restricted, the number of false positive detections could be reduced further.

For further analysis purposes, the 1499 quantified proteins in the AROM+ dataset were subjected to the comparison of differential statistical methods. In this dataset, ROTS and RP produced the lowest number of differentially expressed proteins, whereas SAM provided the highest number. Additionally, 83 differentially expressed proteins were found by all five methods. (FDR<0.05; Figure 10). Of the methods, ROTS provided good agreement with other methods and did not produce false positive detections when biological replicates were compared to each other. Accordingly, using the ROTS method (FDR<0.05), 166 proteins demonstrate differential expression between AROM+ and WT.

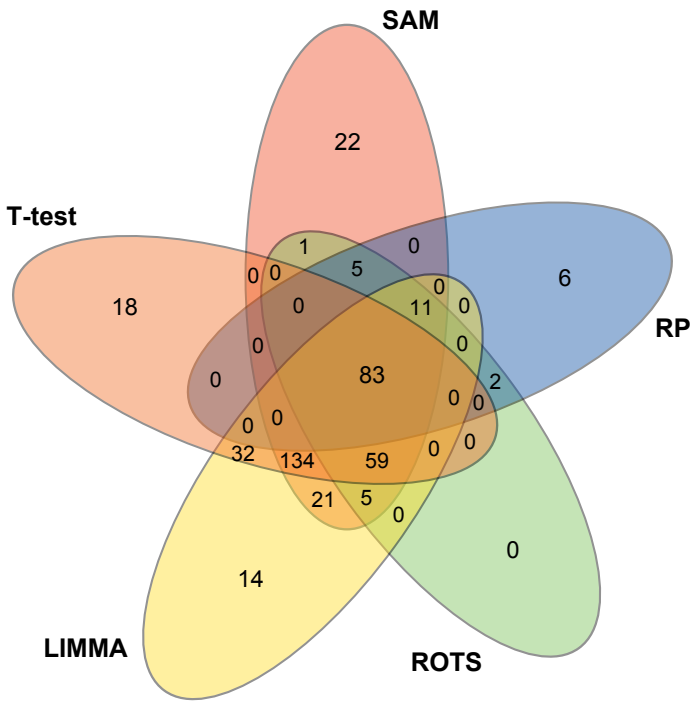

Figure 10. The numbers of differentially expressed proteins in AROM+ dataset detected and shared by different statistical methods.

5.3 Sex hormones affect liver lipid metabolism and plasma metabolite profiles in AROM+ mice (II)

In the aromatase mouse model (AROM+), the human aromatase enzyme is universally expressed resulting in increased estradiol and reduced testosterone production in adult male mice. According to the GC-MS/MS serum steroid screening that was performed to the mouse strain, it was concluded that the AROM+ male mice have a significantly lower testosterone and higher estradiol levels in the serum as compared with their WT littermates (Table 3). To elucidate the effect of the changed male/female sex hormone balance in the adult male AROM+ mouse, an integrated approach was used, where proteomics and transcriptomics data from mouse liver was studied together with the plasma metabolite profiles. As a result, a clear effect of the increased circulating estradiol to testosterone ratio was found in both the liver physiology and plasma metabolites, observed as a distinct clustering of AROM+ and WT based on all three omics datasets.

Table 3. The results of the GC-MS/MS serum steroid analyses performed on male AROM+ and WT mice. Used with permission from L. Strauss.

Hormone	WT	AROM+
Estradiol, pg/mL	< 0.3[b]	30.9 ± 18.4
Testosterone, pg/mL	4972 ± 2067[a]	385 ± 58
Estrone, pg/mL	< 0.5	9.7 ± 0.7
Androstenedione, pg/mL	180 ± 67	165 ± 21

Analyzed with Mann-Whitney Rank Sum Test Values are given as mean ± SEM. [a]P < 0.01, [b]P < 0.001 vs AROM+ mice.

When the AROM+ mouse liver tissues were compared to the WT, 159 differentially expressed mRNAs were found in the transcriptomics experiments and 166 differentially expressed proteins were identified using LC-MS label free quantification and the ROTS method. The human aromatase P450 was not detected with either one of the approaches, as the microarray chip (MouseWG-6 v2.0; Illumina, Essex, UK) and the Uniprot protein sequence database were both based on murine sequences. An integrative analysis scheme of the transcriptomics and proteomics data was chosen, although they were expectedly poorly correlated (r^2=0.54). Consequently, the two datasets were subjected to Ingenuity pathway analysis (IPA, www.ingenuity.com) to find pathways in which the differentially expressed mRNAs or proteins were enriched. As a result, the pathways

of fatty acid metabolism, metabolism of xenobiotic and metabolism of terpenoid were found significantly enriched in both datasets. Moreover, the pathway analysis suggested that multiple nuclear receptors related to estrogen function are upstream regulators of the differentially expressed mRNAs and proteins. These include retinoic acid receptor-related orphan receptor alpha (RORα) and peroxisome proliferator-activated receptor alpha (PPARα).

Furthermore, the mRNAs and proteins differentially expressed in these networks were often related to peroxisomal ß-oxidation. Interestingly, a dual effect on the fatty acid oxidation was found. In the classical peroxisomal ß-oxidation pathway multiple upregulated proteins and mRNAs confirmed in qRT-PCR experiments, such as Ehhadh, Acaa1b and Acot3 were implicating an accelerated oxidation of long fatty acids in the liver of AROM+ mice. Conversely, the oxidation of branched fatty acids, fatty dicarboxylic acids and ω-oxidation was found downregulated, exemplified by a dramatic downregulation of mRNAs and proteins for Cyp4a12a and Scp2 by qRT-PCR. Interestingly, also a group of gender specific mRNAs and proteins were deregulated in AROM+ male mice indicating a feminization of liver metabolism in the male mice.

Importantly, metabolites studied in the plasma of AROM+ male mice reflect these changes. A quantitative screen of 163 metabolites in mouse plasma, including free carnitine, acylcarnitines, amino acids, hexoses and glycerophospholipids resulted in the identification of 62 differentially expressed metabolites between AROM+ and WT male mice. Among these metabolites, especially plasma phospholipids with long fatty acid acyl tails were found significantly decreased in the plasma of male AROM+ mice, whereas phospholipids with an arachidonic acid (C20:4) tails were found increased. Furthermore, when the data from female mice was integrated to the analysis, a feminization of the plasma metabolite pattern of AROM+ male mice is visible (Figure 11). In summary, we identified mRNAs and proteins responsive to imbalanced testosterone-estradiol ratio in the liver and a systems wide effect of plasma phospholipids in male mice and while there may not be a measurable correlation between the data types, there is a clear trace of molecular activity.

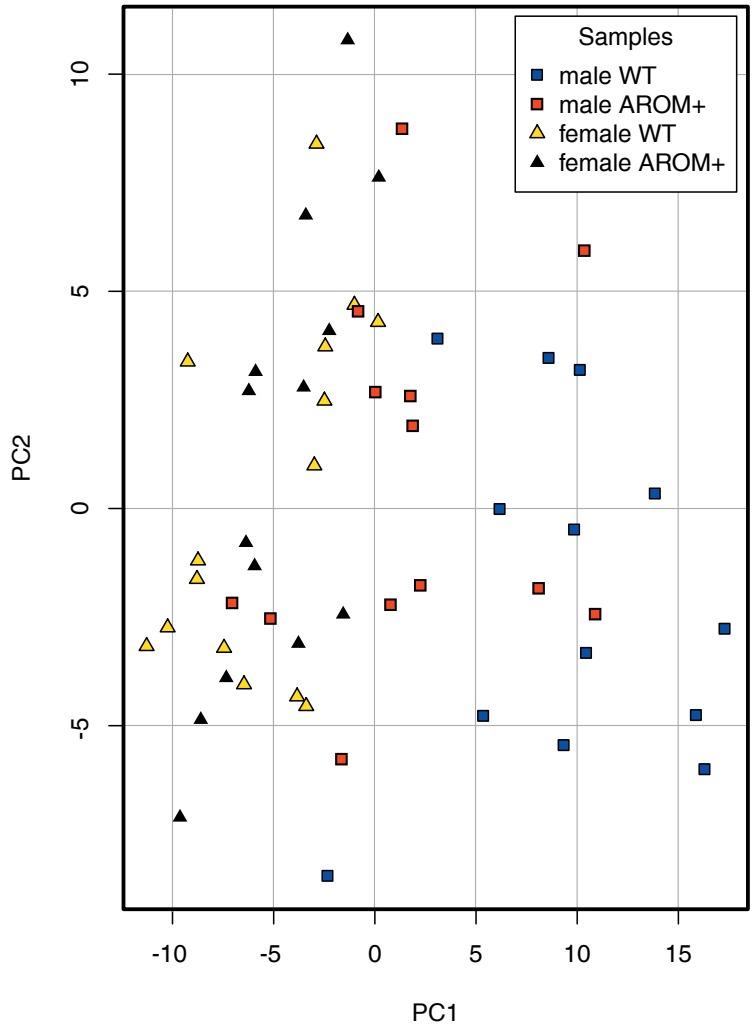

Figure 11. PCA analysis of 162 metabolites screened in the plasma of AROM+ and WT male and female mice.

5.4 Biomarker candidates identified in ovarian endometrioma (III)

In the endometriosis study, the label-free quantification workflow was applied on peptide samples directly derived from miniscule amounts of patient endometrioma and endometrium tissues. The methods optimized and applied during this study allowed the identification and quantification of putative markers for ovarian endometriosis, which are now being validated in a follow-up study. Overall 214 differentially expressed proteins were identified in ovarian endometrioma compared to endometrium. Based on these results, a list of transitions for subsequent SRM verification work was constructed and made publically available. Moreover, a further pathway analysis was accomplished

in IPA. As a result, multiple significantly enriched pathways were found in the proteomics data and among the most interesting pathways identified was the transforming growth factor beta-1 (TGF-β1) -dependent signaling. This pathway is of importance as one third of the differentially expressed proteins in ovarian endometrioma were found to be under TGF-β1 regulation, according to the pathway analysis.

Similarly to the mouse liver samples, the proteomics results were integrated to mRNA microarray analyses performed on ovarian endometrioma and patient endometrium. In this manner, the most potent candidates were identified and the numbers of putative markers could be controlled. As a result, for 88 of the differently expressed proteins, a significantly different expression in transcript level was found, including nine of the biomarker candidate proteins targeted with the AIMS experiment. Thus, as an outcome of the mRNA-protein comparison, a compressed list of 41 upregulated putative biomarkers was constructed.

To pinpoint the specificity of these proteins to ovarian endometrioma and to exclude the effect of ovarian stroma possibly present in samples, an in silico specificity study was conducted. In the study, the expression levels of biomarker candidates were evaluated using databases for tissue-level expression of proteins in uterus, ovary and ovarian endometrioma. The highly ranked biomarker candidates showed high specificity of endometrium and ovarian endometrioma but had only low expression levels in ovary. As an outcome, a list of 14 TGF-β1 regulated, transcriptionally upregulated, highly specific biomarker candidate proteins for ovarian endometrioma was constructed. Moreover, the localization and upregulation of two candidates on this list, Calponin-1 and EMILIN-1, were verified by immunohistochemistry and Western blot analyses on ovarian endometrioma.

6. DISCUSSION

During the work it became evident that the complications in finding a biomarker using proteomics and mass spectrometry rise from multiple sources within the full peptide identification and quantification workflow. Therefore, due to time, resource and technical constraints, compromises between the general robustness of the quantitative analysis and proteome coverage were made throughout the study. As a result, using a robust analysis, a good coverage of the proteome of the liver and endometriosis was reached.

6.1 Technical aspects related to identification and quantification (II, III)

The samples used for this work present the two opposites of tissue samples. The work with ovarian endometrioma tissue samples offers challenges typical for a clinical patient samples: They are inherently non-renewable, unique and limited. More specifically, only small amounts of samples were available and the patients had to be matched according to their age, BMI and menstrual cycle phase. Also, in addition to problems associated with the heterogeneous nature of endometriotic tissue itself (as discussed in chapter 2.3.2), a possibility existed that the ovarian endometrioma samples were contaminated by cells of ovarian origin. Conversely, the mouse liver tissue is often used as a model sample for in-depth proteome explorations (as discussed in chapter 2.3.2). The model system used, P450 aromatase over-expressing mouse (AROM+), provided with low inter-mouse variance and proved to be an excellent model to study proteome level changes in the liver after constant over-expression of an enzyme affecting the liver function.

The analytical differences between the mouse liver and endometriosis samples were clearly demonstrated in the data dependent (DDA) analysis. In these results, the number of identified peptides was found to be lower in each individual endometriosis sample (3700 sequences) when compared to AROM+ liver samples (8500 sequences). It was realized, however, that the difference was substantially reduced when the entire endometriosis and AROM+ datasets were studied (10794/12284 sequences; as discussed in chapter 5.1). These differences illustrate the possibly lower dynamic range of protein concentrations in the mouse liver sample, when compared to endometriosis, which makes the sample better suited for DDA analysis. However, as different preparation methods were used for these samples, the differences may be also due

to a more efficient sample extraction and separation techniques chosen for AROM+. Nevertheless, the combination of data from multiple samples increased the relative number of identifications acquired from endometriosis tissues more than it did in the mouse liver samples. This is likely due to a larger molecular variability of endometriosis tissues. Indeed, higher inter-sample variability was also measured for this sample type during the label-free quantification analysis (see chapter 5.1).

As sample preparation is one of the largest sources of variability in any quantitative proteomics experiment (Krüger et al, 2013; Piehowski et al, 2013) and the label-free quantification is performed without labels that could be used to normalize the sample amounts, the sample preparation in this study was performed with minimum sample handling without fractionation (see chapter 4.1.1 and 5.1). More specifically, this choice was made to avoid the error-prone division of samples to different fractions and the possibly laborious and challenging re-combination of fractions either before or after quantification. Moreover, as labels are not added, the use of a label-free quantification method necessitates the optimization of the chromatographic set-up for robustness.

In this study, one of the major challenges in the reverse phase chromatographic separation was to maintain an adequate precision for suitable quantification. An acceptable sample-to-sample repeatability was accomplished by the use of splitless nanoliter flow rate LC system (Easy nanoLC) with a vented column configuration. In this system the pre-column resides next to the analytical column and provides minimum inter-column dead volumes (Licklider et al, 2002). Moreover, to increase robustness, the sample injection amount was standardized, the samples were injected in a random order and a short wash run was performed between the samples. To monitor the performance of the system, the ion intensities of different samples were observed by overlaying the total ion chromatograms and spray stability plots of different injections in real time. Also repeated injections of complex samples were performed to monitor repeatability. As a result the coefficient of variance was found to be 22 % for the endometriosis and 13 % for liver dataset prior normalization (See chapter 5.1). A likely explanation for the lower technical variance for the liver samples could be the quality of the sample: The liver proteins were precipitated with acetone prior digestion and the resulting peptides were subjected to C18 sample clean up prior injection. Nevertheless, similar numbers for technical variation have been reported for brain tissue (Piehowski et al, 2013), urine (Nagaraj & Mann, 2011) and plasma (Qian et al, 2009) , and while the instrumental variance depends on a sample, the result is acceptable.

The key issue of setting up the LTQ Orbitrap Velos instruments for the label-free experiments was to ensure enough measurements per precursor ion (MS1) and at the same time maximize the amount of fragment ions (MS/MS). The number of MS1 measurements is important for the right definition of feature intensity, which is essential for accurate quantification (Radulovic et al, 2004), whereas the number of MS/MS cycles is directly proportional to the number of identified peptides. Thus, to reach optimal conditions in data dependent mode, the duty cycle of the instrument has to be as fast as possible. In this study, the settings used for MS and MS/MS in the LTQ Orbitrap Velos systems represented a compromise between sequencing speed and quality of quantification. The number of MS/MS events per cycle in LTQ was reduced from the maximum 20 to 15 and the Orbitrap analyzer was operated with a resolution of 60 000 instead of maximum resolution of 100 000 (See chapter 4.1.2). Even though the maximum resolution would provide more accurate quantification and offer better separation of features in the LC-MS map, it extends the duty cycle of the instrument to two seconds. This is too slow for the detection of optimal number of spectra over an eluting peak and to provide a consistent MS1 pattern for the quantification algorithm to perform reliable feature detection. Nevertheless, despite the attempts to optimize the chromatographic separation and the instrumental duty cycle, it is clear that by analyzing the samples without fractionation certain lower abundant peptides and co-eluting species will be missed in the sample (Geiger et al, 2012).

To further increase the proteome coverage in the endometriosis and liver datasets, two different inclusion approaches were exploited: 1) Accurate inclusion mass screening (AIMS; Jaffe et al, 2008) and 2) a general inclusion method (directed MS; Schmidt et al, 2008). With AIMS specific proteins are targeted in the analysis by programming the instrument to analyze proteotypic peptide masses specific for the protein in question. This approach was used in the endometriosis datasets for a set of 18 previously defined endometriosis biomarker candidates, which were expected to be found in the sample based on a preliminary mass spectrometry and microarray analysis. When the 132 proteotypic peptide masses of the 18 proteins were analyzed in the endometriosis samples, a 19 % increase in identifications was reached as compared to the analysis without inclusion list (See chapter 5.1). Additionally, a general inclusion method (Schmidt et al, 2008) was applied in both liver and endometriosis samples to those precursors that were detected but not identified in previous analyses of the same sample.

The inclusion method was restricted to all detected but unidentified MS1 spectra in the individual endometriosis samples and to those detected MS1 spectra that remained unidentified or were fragmented but identified with only one PSMs in the mixed liver sample. Interestingly, this difference in the approach is visible in the results: In individual samples the increase in unique identified sequences was on average 8 % in endometriosis for each consecutive run, whereas in the liver sample the number is only 4 % (See chapter 5.1). However, again, when the increase in unique sequences is studied at the level of the entire dataset, the numbers are reversed: The general increase of identified sequences was recorded to be 12 % in endometriosis and as high as 30 % in the liver samples (See chapter 5.1). This inconsistency could be due to the higher relative heterogeneity of endometriosis tissues when compared to liver. Another cause could be the use of more efficient inclusion strategy in the liver sample counting also those uncertain identifications with one PSM in the inclusion experiment. This strategy allows the instrument to focus on those peptides that are identifiable, but present in only small quantities. Importantly, the mixed liver sample was subjected to four rounds of consecutive inclusion of unidentified precursor ions, whereas in the endometriosis study each sample was analyzed only two times. The number of inclusion rounds has previously been reported to impact the number of identified peptides substantially (Schmidt et al, 2008).

To acquire protein identifications, a database search of spectral data was conducted (See chapter 4.1.3). This experiment was also designed to reach a compromise between a robust set of identifications and an adequate coverage. Non-redundant, fully annotated Uniprot-Swiss-Prot databases for human and mouse were used to minimize the effect of shared peptides between protein isoforms and to reduce the number of sequences subjected to the search. This was accomplished together with a shuffled decoy database search to control the influence of false positive identifications. To optimize coverage, all spectral files in each experiment were searched together. This combined approach is logical in LC-MS label-free quantification experiment, where the data files are analyzed together, and increases the likelihood of receiving an acceptable score of multiple low scoring spectra originating from different spectral files. Moreover, the spectra were searched by two database software packages, Mascot and Sequest, simultaneously, an approach which has shown to increase the number of identified spectra up to 38 % (Shteynberg 2013). Finally, for Progenesis quantification software, a prerequisite for a protein was to have at least two spectra that were matched to peptides (PSMs) and at least one peptide having a sequence specific for the protein (unique peptide).

The LC-MS label-free quantification step was also optimized for robustness. This was performed by a careful inspection of the LC-MS map for chromatographic tailing, contaminations, uneven loading or ionization suppression. After feature detection, the general effect of normalization was studied by calculating the median CV% of all non-normalized and normalized features. Reduced variance is expected after successful normalization, which was true in our data: normalization reduced the variability from 22 to 10 % in the endometriosis samples and from 13 to 10 % in the liver dataset (See chapter 5.1). The feature maps were also inspected for overlapping and split features. It was observed that when the features with short elution times (<12 seconds) were filtered out of the datasets, the number of split features could be substantially reduced. This filtering was performed to increase the robustness of the quantification, even though high peak capacity is considered to be advantageous in reaching high coverage of the proteome (Köcher et al, 2011b). Furthermore, strict criteria were applied to lists of identified peptides that were imported into the quantification software. To reduce ambiguity, the identifications were searched and imported as a single file. Further refinement included the requirement of at least two spectra per feature with a mass tolerance of lower than 5 ppm. In this manner, those features associated with one spectrum only, as well as those precursors with incorrectly assigned second isotopes were filtered out (Sandin et al, 2015). As a consequence, the application of these filtering steps undoubtedly prohibited the quantification of certain low abundant peptides. As a result, 81 % of the identified proteins in the AROM+ dataset and 72 % of the identified proteins in the endometriosis dataset were subjected to quantitative analysis.

6.2 Optimization of Statistical Methods for Quantitative Proteomics (I)

Statistical evaluation of proteomics data can be compared to that of mRNA microarrays: the data matrix is complex and the number of measurement points immensely exceeds the number of samples (Christin et al, 2013). Therefore, it is crucial to correct for multiple testing. Also the variability between samples can be high and the need for optimized normalization is of vital importance. In this study, five different algorithms originally developed for microarray data analysis were evaluated for LC-MS label-free quantitative data.

In the experiment, five different statistical approaches were evaluated in a dataset, where five different concentrations of mixture of 48 human proteins were spiked to

a complex yeast protein background (See chapter 5.2). While all the methods were found suitable for proteome data analysis, methodological differences between the approaches were found. All methods performed poorly in the lowest and highest concentration range of spiked proteins. The deficient quantification of lower abundant proteins (2 and 4 fmol/µl) is likely to be a result of reaching the lower limit of detection for this instrument, whereas the in the higher range of spiked proteins (25 and 50 fmol/µl) the problems in quantification may be caused by matrix interference and saturation effects. However, in a mixture of proteins where the concentrations are ranging from 13 ng/µl (Epidermal Growth Factor; 2 fmol/µl) to 4150 ng/µl (Gelsolin; 50 fmol/ µl), some species clearly performed better than others in high and low concentrations. In addition, all methods provided reasonably high numbers of false positive identifications (i.e., significantly changed yeast proteins). This phenomenon might be due to intrinsic effects in LC-MS/MS, such as matrix interference, but can be also some other artefact for technical variability, as lower numbers of false positive identifications were received after optimized normalization. Moreover, as reported previously (Gregori et al, 2013), the numbers of false positive detections could be reduced considerably with the application of a fold change threshold for a significantly changed protein and by increasing the number of required peptides per protein to two.

When tested on the AROM+ mouse liver dataset, different methods provided very different results in terms in the numbers of significantly changed proteins, ranging from 359 detected by LIMMA to 107 proteins detected by Rank Product method. The result is problematic, as the different numbers of detections can be a result of either false positive or negative detections. Indeed, variable amounts of false positive identifications were detected when biological replicates were compared to each other. Especially Significance Analysis of Microarrays (SAM) provided high rate of false positive identifications. Moreover, based on the results acquired in the yeast dataset, higher false positive rate is expected for SAM and, conversely, high false negative rate for Rank Product. Throughout the tests, the Reproducibility-Optimized Test Statistic (ROTS) and Linear Models for Microarray Data (LIMMA) presented the best sensitivity and specificity, while the number of false positives stayed in mid-range when compared to the other statistical tests evaluated. Furthermore, when compared to other methods in the AROM+ dataset, ROTS showed good agreement with the results of other tests applied.

6.3 Overexpression of aromatase enzyme alters liver lipid metabolism in male mice (II)

For the purpose of receiving an overview on the metabolic events in the liver of AROM+ male mice, the transcriptome, proteome and metabolome datasets were integrated using a multi-stage approach, where all datasets are first analyzed independently and subjected to subsequent comparison (See chapter 2.3.3). This approach relies on the assumption that proteome and metabolome level changes are triggered by changes in the gene expression rather than on a level of a proteome (Ritchie et al, 2015). Moreover, as the correlation of mRNA and protein data was expectedly modest (See chapter 5.3), this approach was extended to pathway analysis identifying common patterns of metabolic regulation in the transcriptome and the proteome datasets. The pathway analysis revealed a clear enrichment of fatty acid metabolism and regulators related to estrogen function, such as PPARα, and a substantial upregulation of mRNAs and proteins related to PPARα regulated peroxisomal β-oxidation. However, the most highly deregulated mRNA, Cyp4a12a, showed a dramatic decrease both at transcriptional and protein levels. This particular gene has been found to be responsible of specific ω-oxidation hydroxylation reaction of arachidonic acid to an eicosanoid and the corresponding dicarboxylic acid (Arnold et al, 2010; Muller et al, 2007). These changes in peroxisomal β- and ω-oxidation in the liver reflect the plasma phospholipid profile, where a distinct reduction in the levels of long chain fatty acids could be detected in male AROM+ mice. Moreover, increased amounts of arachidonic acid was discovered, proposing pro-inflammatory features in AROM+ male mice (Harizi et al, 2008).

The association of estradiol and PPARα is known from multiple studies (Barros & Gustafsson, 2011; Zhu et al, 2013). Furthermore, RORα, which is regulated by estradiol and testosterone in neuronal cells, has been reported to induce the expression of PPARα in the liver (Lu et al, 2014; Sarachana & Hu, 2013). The inductive effect of estradiol has also been demonstrated in context of β- and ω-oxidation in liver (Fernández-Pérez et al, 2014). Moreover, an opposite effect has been previously detected in aromatase knock out mouse model (Nemoto et al, 2000). The activation of PPARα is undoubtedly not the only effect of increased estradiol and decreased testosterone levels in AROM+ male mouse liver. It has been shown that estrogen receptor alpha (ERα) regulates glucose homeostasis and lipogenic gene expression in the liver, independent of gender (Bryzgalova et al, 2006). Indeed, we found multiple gender predominant mRNAs and proteins expressed differentially between AROM+ and WT male mice, including reduced levels of the male specific Cyp4a12a. Interestingly, the plasma metabolites analyzed in

both female and male AROM+ and WT mice reflect the partial feminization of the liver surprisingly well. In an unsupervised PCA plot of all metabolites measured in both male and female mice, a clear feminization effect for AROM+ male mice can be seen (See chapter 5.3).

6.4 Identification of specific TGF-β1 regulated markers for ovarian endometrioma (III)

In the endometriosis study, using LC-MS label-free quantification, over 200 proteins were identified as differentially expressed between ovarian endometrioma and patient endometrium tissue. Over 80 % of these proteins have not been previously reported in ovarian endometriomas, which, together with the preliminary validation studies, demonstrates the usefulness of the technology in the work described here.

Similar to the mouse liver work, an integrative approach combining multiple datasets was applied in the further data analysis of the endometriosis dataset. The pathway analyses indicated multiple deregulated pathways in ovarian endometrioma and revealed that a large part of deregulated proteins in the ovarian endometriomas are under the control of TGF-β1 growth factor (See chapter 5.4). Multiple groups have reported the involvement of TGF-β1 in endometriosis (Barcena de Arellano et al, 2011; Sohler et al, 2013; van Kaam et al, 2008; Young et al, 2014). There have been proposals of the role of TGF-β1 in endometriosis in smooth muscle metaplasia (Fukunaga, 2000; Gabbiani, 2003), stromal decidualization (Kim et al, 2005; van Kaam et al, 2008) or accumulation of lactate inducing cell invasion, angiogenesis, and immune suppression (Young et al, 2014). Interestingly, TGF-β1 has been found to induce epithelial mesenchymal transition (EMT), which has been suggested to be part of the cause of endometriosis and is characterized by imbalance of E-cadherin and N-cadherin expression in the epithelial cells (Bartley et al, 2014). However, even though we found a significant increase of N-cadherin mRNA in ovarian endometrioma samples, our proteomics data does not fully support the EMT model. Nevertheless, these results highlight that multiple molecular events exist in ovarian endometriomas that may contribute to a better definition of endometriosis disease in the future.

As the results from the discovery experiment using LC-MS label-free quantification were promising, the attempts to identify a biomarker for ovarian endometriosis were continued by a multi-stage integrative analysis. This included a comparative study of the LC-MS label-free quantitative proteomics results and mRNA data collected from

a larger set of corresponding ovarian endometrioma and endometrium tissues. The comparative study aimed at the qualification of biomarkers candidates to identify a smaller set of relevant candidates for the targeted verification step in the biomarker discovery pipeline (See chapter 2.3.1.). The intention was to identify the differentially expressed proteins and mRNAs in both datasets as key molecules possibly involved in the pathogenesis of ovarian endometrioma. After the comparison, a reduced list of biomarker candidates that were found upregulated in both protein and transcript level, was collated. A qualitative analysis was performed to define the specificity of these proteins and as a result a TGF-β1 regulated sublist of 14 proteins with high specificity for ovarian endometrioma was made available. Interestingly, only recently, five of these proteins were found modulated in ectopic endometriotic stromal cells (Kasvandik et al, 2015).

7. SUMMARY

It is well known that cell-lines undergo changes in gene expression and PTM patterns during cell culturing (Derda et al, 2009) and recently the US National Cancer Institute (NCI) has replaced its cell-line based cancer models with those originating from clinical samples. Indeed, patient tissues, despite their intrinsic heterogeneity of biomolecules and differences from patient to patient, retain the organization of specialized cells and the intracellular communication patterns as well as the nutrients, metabolites and signaling molecules *ex vivo*. However, the complexity of tissues as a sample material for proteome interrogation has prohibited their use until recently. Advanced proteomics methodologies now allow both deeper insight into the tissue proteome and more reliable measurements allowing multi-patient studies, which has been at the foundation of the work pursued here.

In this thesis, analyses of tissue samples with minimal sample handling and fractionation were accomplished by LC-MS label-free quantification. These methods require less sample, sample preparation steps and MS measurement time than methods relying on label-based quantification combined with extensive fractionation. The LC-MS label-free quantitative methods are also generic and thus suitable for any sample type. In addition, the approach has a vast potential for further technical development as it has been estimated that over 100 000 peptides are eluted during a single LC run (Michalski et al, 2011a). Whereas current MS instruments are only able to analyze a small fraction of these peptides, modern high-speed MS instruments will enable the analysis of higher shares of these peptides in the future (Eliuk & Makarov, 2015). However, the LC-MS label free quantification technology has also disadvantages, some of which have been discussed in this thesis. The data-dependent sampling of precursor ions results in incomplete and partially irreproducible numbers of identifications, reflected by variable results from technical replicate analyses. This problem is highlighted in the analysis of tissue samples where the complex MS spectra result in relatively low proteome coverage, as only the most abundant precursor ions are fragmented. In comparison to label-based methods, label-free analyses are also more prone to technical bias that affects the robustness and accuracy of the method as we and others (Russell & Lilley, 2012) have found. It has also been noted that the choice of the quantification software (Sandin et al, 2011) and statistical methods used, can have a large impact on the results of LC-MS label-free quantification.

Optimization of the statistical approach to a high dimensional quantitative proteomics data is imperative and follows the requirements of gene expression microarrays (Boulesteix & Sauerbrei, 2011; Clarke et al, 2008). In the first part of the thesis, five different statistical methods were evaluated for our LC-MS label-free quantitative approach. Concerning the choice of the best statistical method, a number of key issues were identified. Nearly all algorithms offered high sensitivity and specificity for proteins and peptides of medium abundance, but did not perform that well with higher and lower concentrations of the analytes. In addition, high numbers of false positive identifications were not uncommon among the different algorithms tested. However, after optimized normalization and stricter requirements for significance, in terms of minimum number of peptides and fold change of a protein, the number of false positive detections could be reduced. Also, when tested in all possible combinations of WT AROM+ liver samples alone, it was discovered that some algorithms provided higher numbers of false positive or false negative detections, reflected in very different numbers of proteins with significantly changed expression. In conclusion, the in-house developed ROTS method proved to be flexible and robust when tested with both spiked yeast samples and AROM+ mouse liver datasets. This study highlights the importance of the optimized statistical approach for a large scale proteomics datasets and encourages the proteomics community to adapt tools originally developed to gene expression microarrays for the quantitative proteomics data. The standardization of statistical approaches would enhance the reproducibility across different laboratories and raise awareness of the need of validated statistics for proteomics experiments.

In the second and third study, a robust data collection platform was established to AROM+ mouse liver and endometriosis tissue samples allowing data collection with a technical coefficient of variance of 10 %. To avoid fractionation and to reduce technical bias, a DDA data collection was combined with a directed MS approach, where low abundant peptides were targeted with inclusion lists. This approach was successfully used to increase the proteome coverage of both AROM+ mouse liver and endometriosis tissues. Interestingly, approximately 1500 proteins were quantified in both datasets, which is in line with previous reports of LC-MS label-free quantification performed on fractionated samples (Bracht et al, 2014; Vasilj et al, 2012). In this work only non-modified peptides were analyzed due to technical restrictions, even though PTMs, such as phosphorylation and glycosylation, are extremely important modulators of biological effects. Nevertheless, even though the proteomes described here are far from complete and do not challenge the in-depth tissue screens reported recently (Azimifar et al, 2014),

they are quantitative and exceed ten times the quantitative depth previously described for tissue-based proteomics in endometriosis, for example (Hwang et al, 2013).

Multiple diseases, such as type 2 diabetes, hypertension, stroke and cardiovascular disease, have differential prevalence among men and women. It is also known that women have less atherogenic blood lipid profile and differential body fat distribution to men (Palmer & Clegg, 2015). Also gender differences in plasma metabolites have been reported lately (Krumsiek et al, 2015; Mittelstrass et al, 2011). These sexual dimorphisms are known to be estrogen related, but the mechanisms of estrogen action are poorly understood. The altered lipid metabolism we found in AROM+ male mouse liver highlights the role of the organ in mediating the estrogen actions. When studied together with other large scale datasets, such as mRNA microarrays and metabolomics, information about the biological state of AROM+ liver could be extracted from the proteomics data. As a result, a novel link between the altered estrogen-androgen ratio, liver lipid oxidation and lipid metabolites in AROM+ plasma was found, suggesting a reduced oxidation of long chain fatty acids and increased abundance of pro-inflammatory lipids in the liver of AROM+ male mice. As our results so far can be considered a pure discovery study, an important next step in the project would be the gathering of more mechanistic data on the impact of estrogen to androgen ratio to lipid metabolism. This might help us to understand the differential prevalence of common diseases such as atherosclerosis and offer more tailored treatment options in the future.

Despite a clear need for endometriosis related biomarkers, no clinically validated biomarker exists (Fassbender et al, 2015). This is important, as endometriosis burdens the healthcare systems with direct and indirect costs ranging from 4 400 to 28 000 $ per patient yearly worldwide (Soliman et al, 2016). The disease is associated with anxiety and depression, diminishing the quality of life of the patients (Friedl et al, 2015). Especially, pelvic pain, delayed diagnosis and infertility have recently reported to significantly affect the well-being of endometriosis patients (Facchin et al, 2015; Hamdan et al, 2015; Nnoaham et al, 2011). In this study, TGF-β1 was identified as a key regulatory molecule in ovarian endometriosis with the LC-MS label-free quantification method. Moreover, a number of proteins highly specific for ovarian endometrioma were discovered. However, the findings of this study are an outcome of a discovery experiment requiring follow-up experiments for possible biomarker validation. For this reason, the proteins identified as endometrioma specific as well as other published biomarker candidates should be exposed to a full biomarker validation pipeline, starting

with verification, validation and clinical validation steps. Therefore, a high throughput SRM-based biomarker identification and validation in matched plasma samples of endometriosis patients and healthy controls is already underway, our hopes being able to identify markers predicting the fertility of the patient and/or progression of the endometriosis disease.

ACKNOWLEDGEMENTS

This work was carried out at the Turku Centre for Biotechnology and at the Department of Physiology at the University of Turku. I am very grateful for being able to work in such excellent research facilities and inspiring atmosphere. I would also like to acknowledge Professor Ludwig Wildt for the great hospitality at the department of Gynecologic Endocrinology and Reproductive Medicine at the Medical University of Innsbruck during the last two years of the thesis project.

I would like to express my gratitude to my supervisor Professor Garry Corthals for introducing me to the captivating world of proteomics and mass spectrometry and Professor Matti Poutanen for always pushing me forward and offering me generously new opportunities when urgently needed. Furthermore I would like to thank Dr. Kaisa Huhtinen for guiding me through a large clinical project, where the number of samples is certainly much more than the sum of the patients. I would also like to warmly thank Dr. Leena Strauss for the expertise with the AROM+ mice and for the help as well as the positive and supporting attitude at the last phases of the PhD project. This work would not have been possible without the guidance and help by Garry, Matti, Kaisa and Leena and I am deeply grateful for that.

I would also like to thank Professor Per Andrén for his support on mass spectrometry and peptide analytics, the constructive comments during thesis committee meetings and the enthusiasm about my research project. Dr. Hannu Koistinen, DSc. and Dr. Charles Pineau are warmly thanked for reviewing this thesis and providing constructive comments.

I would also like to acknowledge the Turku Doctoral Programme of Molecular Medicine, TuDMM and the National Doctoral Program in Informational and Structural Biology, ISB and its director Professor Mark Johnson for providing support and organizing great social events, such as the legendary Winter Schools in Lapland.

This project would have felt very long without my colleagues Anne, Petri, Susumu, Arttu, Veronika, Susanne and Pekka at proteomics research group and all others I had pleasure to work with. Thank you all for your friendship, help, advice, and sharing the good and bad moments. I would like to especially thank Dr. Beata Seeber and Dr. Fabian Lunger for the very welcoming atmosphere at the Medical University of Innsbruck and Dr. Milla-Riina Neffling for sharing the enthusiasm for mass spectrometry.

I am indebted to my co-authors, whose help made this work possible. I would especially like to thank Dr. Dorota Muth for her help and knowledge in SRM and Dr. Laura Elo, Dr. Gabi Kastenmüller and Daniel Laajala for the bioinformatics collaboration.

I owe special thanks to my family and friends. Without the continuous support from you I would not have reached this point. I also owe thanks to my family in law, who have made this book possible by offering help selflessly whenever needed. Johannes, thank you for your love, encouragement and support through difficult times and in general over the years. Lauri, thank you for showing me the bigger picture – every day.

This work was financially supported by the Academy of Finland, Sigrid Juselius foundation, Finnish Cancer foundation, University of Turku, Orion foundation and by travel grants from the University of Turku foundation, Emil Aaltonen foundation, Systems Biology Research Program and National Doctoral Program in Informational and Structural Biology.

REFERENCES

(1997) Revised American Society for Reproductive Medicine classification of endometriosis: 1996. *Fertil Steril* **67:** 817-821

Abraham JE, Maranian MJ, Spiteri I, Russell R, Ingle S, Luccarini C, Earl HM, Pharoah PP, Dunning AM, Caldas C (2012) Saliva samples are a viable alternative to blood samples as a source of DNA for high throughput genotyping. *BMC Med Genomics* **5:** 19

Abruzzo PM, Ghezzo A, Bolotta A, Ferreri C, Minguzzi R, Vignini A, Visconti P, Marini M (2015) Perspective Biological Markers for Autism Spectrum Disorders: Advantages of the Use of Receiver Operating Characteristic Curves in Evaluating Marker Sensitivity and Specificity. *Dis Markers* **2015:** 329607

Adamson GD, Frison L, Lamb EJ (1982) Endometriosis: studies of a method for the design of a surgical staging system. *Fertil Steril* **38:** 659-666

Addona TA, Abbatiello SE, Schilling B, Skates SJ, Mani DR, Bunk DM, Spiegelman CH, Zimmerman LJ, Ham AJ, Keshishian H, Hall SC, Allen S, Blackman RK, Borchers CH, Buck C, Cardasis HL, Cusack MP, Dodder NG, Gibson BW, Held JM et al (2009) Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat Biotechnol* **27:** 633-641

Addona TA, Shi X, Keshishian H, Mani DR, Burgess M, Gillette MA, Clauser KR, Shen D, Lewis GD, Farrell LA, Fifer MA, Sabatine MS, Gerszten RE, Carr SA (2011) A pipeline that integrates the discovery and verification of plasma protein biomarkers reveals candidate markers for cardiovascular disease. *Nat Biotechnol* **29:** 635-643

Ahmed FE (2009) Sample preparation and fractionation for proteome analysis and cancer biomarker discovery by mass spectrometry. *J Sep Sci* **32:** 771-798

Ametzazurra A, Matorras R, García-Velasco JA, Prieto B, Simón L, Martínez A, Nagore D (2009) Endometrial fluid is a specific and non-invasive biological sample for protein biomarker identification in endometriosis. *Hum Reprod* **24:** 954-965

Anderson NL (2005) The roles of multiple proteomic platforms in a pipeline for new diagnostics. *Mol Cell Proteomics* **4:** 1441-1444

Anderson NL, Anderson NG (2002) The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics* **1:** 845-867

Anderson NL, Anderson NG, Haines LR, Hardie DB, Olafson RW, Pearson TW (2004) Mass spectrometric quantitation of peptides and proteins using Stable Isotope Standards and Capture by Anti-Peptide Antibodies (SISCAPA). *J Proteome Res* **3:** 235-244

Andreasson U, Perret-Liaudet A, van Waalwijk van Doorn LJ, Blennow K, Chiasserini D, Engelborghs S, Fladby T, Genc S, Kruse N, Kuiperij HB, Kulic L, Lewczuk P, Mollenhauer B, Mroczko B, Parnetti L, Vanmechelen E, Verbeek MM, Winblad B, Zetterberg H, Koel-Simmelink M et al (2015) A Practical Guide to Immunoassay Method Validation. *Front Neurol* **6:** 179

Annesley TM (2003) Ion suppression in mass spectrometry. *Clin Chem* **49:** 1041-1044

Armbruster DA, Pry T (2008) Limit of blank, limit of detection and limit of quantitation. *Clin Biochem Rev* **29 Suppl 1:** S49-52

Arnold C, Markovic M, Blossey K, Wallukat G, Fischer R, Dechend R, Konkel A, von Schacky C, Luft FC, Muller DN, Rothe M, Schunck WH (2010) Arachidonic acid-metabolizing cytochrome P450 enzymes are targets of {omega}-3 fatty acids. *J Biol Chem* **285:** 32720-32733

Asara JM, Christofk HR, Freimark LM, Cantley LC (2008) A label-free quantification method by MS/MS TIC compared to SILAC and spectral counting in a proteomics screen. *Proteomics* **8:** 994-999

Azimifar SB, Nagaraj N, Cox J, Mann M (2014) Cell-type-resolved quantitative proteomics of murine liver. *Cell Metab* **20:** 1076-1087

Baggerly KA, Morris JS, Coombes KR (2004) Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics* **20:** 777-785

Bailey DJ, McDevitt MT, Westphall MS, Pagliarini DJ, Coon JJ (2014) Intelligent data acquisition blends targeted and discovery methods. *J Proteome Res* **13:** 2152-2161

Barcena de Arellano ML, Gericke J, Reichelt U, Okuducu AF, Ebert AD, Chiantera V, Schneider A, Mechsner S (2011) Immunohistochemical characterization of endometriosis-associated smooth muscle cells in human peritoneal endometriotic lesions. *Hum Reprod* **26:** 2721-2730

Barros RP, Gustafsson J (2011) Estrogen receptors and the metabolic network. *Cell Metab* **14:** 289-299

Bartley J, Jülicher A, Hotz B, Mechsner S, Hotz H (2014) Epithelial to mesenchymal transition (EMT) seems to be regulated differently in endometriosis and the endometrium. *Arch Gynecol Obstet* **289:** 871-881

Bateman NW, Goulding SP, Shulman NJ, Gadok AK, Szumlinski KK, MacCoss MJ, Wu CC (2014) Maximizing peptide identification events in

proteomic workflows using data-dependent acquisition (DDA). *Mol Cell Proteomics* **13:** 329-338

Beck S, Michalski A, Raether O, Lubeck M, Kaspar S, Goedecke N, Baessmann C, Hornburg D, Meier F, Paron I, Kulak NA, Cox J, Mann M (2015) The Impact II, a Very High-Resolution Quadrupole Time-of-Flight Instrument (QTOF) for Deep Shotgun Proteomics. *Mol Cell Proteomics* **14:** 2014-2029

Becker KF (2015) Using tissue samples for proteomic studies-critical considerations. *Proteomics Clin Appl* **9:** 257-267

Bellei E, Bergamini S, Monari E, Fantoni LI, Cuoghi A, Ozben T, Tomasi A (2011) High-abundance proteins depletion for serum proteomic analysis: concomitant removal of non-targeted proteins. *Amino Acids* **40:** 145-156

Bellew M, Coram M, Fitzgibbon M, Igra M, Randolph T, Wang P, May D, Eng J, Fang R, Lin C, Chen J, Goodlett D, Whiteaker J, Paulovich A, McIntosh M (2006) A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics* **22:** 1902-1909

Bereman MS, Johnson R, Bollinger J, Boss Y, Shulman N, MacLean B, Hoofnagle AN, MacCoss MJ (2014) Implementation of statistical process control for proteomic experiments via LC MS/MS. *J Am Soc Mass Spectrom* **25:** 581-587

Berger B, Peng J, Singh M (2013) Computational solutions for omics data. *Nat Rev Genet* **14:** 333-346

Berkley KJ, Rapkin AJ, Papka RE (2005) The pains of endometriosis. *Science* **308:** 1587-1589

Bern M, Finney G, Hoopmann MR, Merrihew G, Toth MJ, MacCoss MJ (2010) Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry. *Anal Chem* **82:** 833-841

Blackburn K, Mbeunkui F, Mitra SK, Mentzel T, Goshe MB (2010) Improving protein and proteome coverage through data-independent multiplexed peptide fragmentation. *J Proteome Res* **9:** 3621-3637

Bondarenko PV, Chelius D, Shaler TA (2002) Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Anal Chem* **74:** 4741-4749

Bonner A, Neupane B, Beyene J (2014) Testing for associations between systolic blood pressure and single-nucleotide polymorphism profiles obtained from sparse principal component analysis. *BMC Proc* **8:** S95

Boschetti E, Righetti PG (2009) The art of observing rare protein species in proteomes with peptide ligand libraries. *Proteomics* **9:** 1492-1510

Boulesteix AL, Sauerbrei W (2011) Added predictive value of high-throughput molecular data to clinical data and its validation. *Brief Bioinform* **12:** 215-229

Bracht T, Hagemann S, Loscha M, Megger DA, Padden J, Eisenacher M, Kuhlmann K, Meyer HE, Baba HA, Sitek B (2014) Proteome analysis of a hepatocyte-specific BIRC5 (survivin)-knockout mouse model during liver regeneration. *J Proteome Res* **13:** 2771-2782

Breker M, Schuldiner M (2014) The emergence of proteome-wide technologies: systematic analysis of proteins comes of age. *Nat Rev Mol Cell Biol* **15:** 453-464

Brusniak MY, Bodenmiller B, Campbell D, Cooke K, Eddes J, Garbutt A, Lau H, Letarte S, Mueller LN, Sharma V, Vitek O, Zhang N, Aebersold R, Watts JD (2008) Corra: Computational framework and tools for LC-MS discovery and targeted mass spectrometry-based proteomics. *BMC Bioinformatics* **9:** 542

Bryzgalova G, Gao H, Ahren B, Zierath JR, Galuska D, Steiler TL, Dahlman-Wright K, Nilsson S, Gustafsson JA, Efendic S, Khan A (2006) Evidence that oestrogen receptor-alpha plays an important role in the regulation of glucose homeostasis in mice: insulin sensitivity in the liver. *Diabetologia* **49:** 588-597

Buermans HP, den Dunnen JT (2014) Next generation sequencing technology: Advances and applications. *Biochim Biophys Acta* **1842:** 1932-1941

Bulun SE (2009) Endometriosis. *N Engl J Med* **360:** 268-279

Cambridge SB, Gnad F, Nguyen C, Bermejo JL, Krüger M, Mann M (2011) Systems-wide proteomic analysis in mammalian cells reveals conserved, functional protein turnover. *J Proteome Res* **10:** 5275-5284

Cao Z, Yende S, Kellum JA, Robinson RA (2013) Additions to the Human Plasma Proteome via a Tandem MARS Depletion iTRAQ-Based Workflow. *Int J Proteomics* **2013:** 654356

Chapman JD, Goodlett DR, Masselon CD (2014) Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. *Mass Spectrom Rev* **33:** 452-470

Chick JM, Kolippakkam D, Nusinow DP, Zhai B, Rad R, Huttlin EL, Gygi SP (2015) A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat Biotechnol* **33:** 743-749

Choi H, Glatter T, Gstaiger M, Nesvizhskii AI (2012) SAINT-MS1: protein-protein interaction scoring using label-free intensity data in affinity purification-mass spectrometry experiments. *J Proteome Res* **11:** 2619-2624

Choi H, Nesvizhskii AI (2008) False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J Proteome Res* **7:** 47-50

Christin C, Hoefsloot HC, Smilde AK, Hoekman B, Suits F, Bischoff R, Horvatovich P (2013) A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. *Mol Cell Proteomics* **12:** 263-276

Cima I, Schiess R, Wild P, Kaelin M, Schüffler P, Lange V, Picotti P, Ossola R, Templeton A, Schubert O, Fuchs T, Leippold T, Wyler S, Zehetner J, Jochum W, Buhmann J, Cerny T, Moch H, Gillessen S, Aebersold R et al (2011) Cancer genetics-guided discovery of serum biomarker signatures for diagnosis and prognosis of prostate cancer. *Proc Natl Acad Sci U S A* **108:** 3342-3347

Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, Wang Y (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat Rev Cancer* **8:** 37-49

Clough T, Key M, Ott I, Ragg S, Schadow G, Vitek O (2009) Protein quantification in label-free LC-MS experiments. *J Proteome Res* **8:** 5275-5284

Colaert N, Gevaert K, Martens L (2011) RIBAR and xRIBAR: Methods for reproducible relative MS/MS-based label-free protein quantification. *J Proteome Res* **10:** 3183-3189

Cooper B, Feng J, Garrett WM (2010) Relative, label-free protein quantitation: spectral counting error statistics from nine replicate MudPIT samples. *J Am Soc Mass Spectrom* **21:** 1534-1546

Corrigan L, Jefferies C, Clive Lee T, Daly J (2011) Evaluation and optimization of IgY spin column technology in the depletion of abundant proteins from human serum. *Proteomics* **11:** 3415-3419

Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26:** 1367-1372

Craig R, Beavis RC (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20:** 1466-1467

Craig R, Cortens JC, Fenyo D, Beavis RC (2006) Using annotated peptide mass spectrum libraries for protein identification. *J Proteome Res* **5:** 1843-1849

Crooke PS, Ritchie MD, Hachey DL, Dawling S, Roodi N, Parl FF (2006) Estrogens, enzyme variants, and breast cancer: a risk model. *Cancer Epidemiol Biomarkers Prev* **15:** 1620-1629

Cutillas PR, Vanhaesebroeck B (2007) Quantitative profile of five murine core proteomes using label-free functional proteomics. *Mol Cell Proteomics* **6:** 1560-1573

Davis MT, Spahr CS, McGinley MD, Robinson JH, Bures EJ, Beierle J, Mort J, Yu W, Luethy R, Patterson SD (2001) Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry. II. Limitations of complex mixture analyses. *Proteomics* **1:** 108-117

de Sousa Abreu R, Penalva LO, Marcotte EM, Vogel C (2009) Global signatures of protein and mRNA expression levels. *Mol Biosyst* **5:** 1512-1526

de Sá Rosa e Silva AC, Rosa e Silva JC, Nogueira AA, Petta CA, Abrão MS, Ferriani RA (2006) The levonorgestrel-releasing intrauterine device reduces CA-125 serum levels in patients with endometriosis. *Fertil Steril* **86:** 742-744

Del Campo M, Jongbloed W, Twaalfhoven HA, Veerhuis R, Blankenstein MA, Teunissen CE (2015) Facilitating the Validation of Novel Protein Biomarkers for Dementia: An Optimal Workflow for the Development of Sandwich Immunoassays. *Front Neurol* **6:** 202

Delmotte N, Lasaosa M, Tholey A, Heinzle E, Huber CG (2007) Two-dimensional reversed-phase x ion-pair reversed-phase HPLC: an alternative approach to high-resolution peptide separation for shotgun proteome analysis. *J Proteome Res* **6:** 4363-4373

Dephoure N, Zhou C, Villén J, Beausoleil SA, Bakalarski CE, Elledge SJ, Gygi SP (2008) A quantitative atlas of mitotic phosphorylation. *Proc Natl Acad Sci U S A* **105:** 10762-10767

Derda R, Laromaine A, Mammoto A, Tang SK, Mammoto T, Ingber DE, Whitesides GM (2009) Paper-supported 3D cell culture for tissue-based bioassays. *Proc Natl Acad Sci U S A* **106:** 18457-18462

Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R (2006) The PeptideAtlas project. *Nucleic Acids Res* **34:** D655-658

DeSouza LV, Taylor AM, Li W, Minkoff MS, Romaschin AD, Colgan TJ, Siu KW (2008) Multiple reaction monitoring of mTRAQ-labeled peptides enables absolute quantification of endogenous levels of a potential cancer marker in cancerous and normal endometrial tissues. *J Proteome Res* **7:** 3525-3534

Di Palma S, Boersema PJ, Heck AJ, Mohammed S (2011) Zwitterionic hydrophilic interaction liquid chromatography (ZIC-HILIC and ZIC-cHILIC) provide high resolution separation and increase sensitivity in proteome analysis. *Anal Chem* **83:** 3440-3447

Diamandis EP (2003) Point: Proteomic patterns in biological fluids: do they represent the future of cancer diagnostics? *Clin Chem* **49:** 1272-1275

Diamandis EP (2004) Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. *J Natl Cancer Inst* **96:** 353-356

Dillen L, Cools W, Vereyken L, Lorreyne W, Huybrechts T, de Vries R, Ghobarah H, Cuyckens F (2012) Comparison of triple quadrupole and high-

resolution TOF-MS for quantification of peptides. *Bioanalysis* **4:** 565-579

Duffy MJ, Sturgeon CM, Sölétormos G, Barak V, Molina R, Hayes DF, Diamandis EP, Bossuyt PM (2015) Validation of new cancer biomarkers: a position statement from the European group on tumor markers. *Clin Chem* **61:** 809-820

Duncan MW, Yergey AL, Patterson SD (2009) Quantifying proteins by mass spectrometry: the selectivity of SRM is only part of the problem. *Proteomics* **9:** 1124-1127

Dutta M, Subramani E, Taunk K, Gajbhiye A, Seal S, Pendharkar N, Dhali S, Ray CD, Lodh I, Chakravarty B, Dasgupta S, Rapole S, Chaudhury K (2015) Investigation of serum proteome alterations in human endometriosis. *J Proteomics* **114:** 182-196

Eeltink S, Dolman S, Swart R, Ursem M, Schoenmakers PJ (2009) Optimizing the peak capacity per unit time in one-dimensional and off-line two-dimensional liquid chromatography for the separation of complex peptide samples. *J Chromatogr A* **1216:** 7368-7374

Egelhofer V, Hoehenwarter W, Lyon D, Weckwerth W, Wienkoop S (2013) Using ProtMAX to create high-mass-accuracy precursor alignments from label-free quantitative mass spectrometry data generated in shotgun proteomics experiments. *Nat Protoc* **8:** 595-601

El-Kasti MM, Wright C, Fye HK, Roseman F, Kessler BM, Becker CM (2011) Urinary peptide profiling identifies a panel of putative biomarkers for diagnosing and staging endometriosis. *Fertil Steril* **95:** 1261-1266.e1261-1266

Eliuk S, Makarov A (2015) Evolution of Orbitrap Mass Spectrometry Instrumentation. *Annu Rev Anal Chem (Palo Alto Calif)* **8:** 61-80

Eng JK, Jahan TA, Hoopmann MR (2013) Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13:** 22-24

Eng JK, McCormack AL, Yates JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **5:** 976-989

Eskenazi B, Warner ML (1997) Epidemiology of endometriosis. *Obstet Gynecol Clin North Am* **24:** 235-258

Etzioni R, Urban N, Ramsey S, McIntosh M, Schwartz S, Reid B, Radich J, Anderson G, Hartwell L (2003) The case for early detection. *Nat Rev Cancer* **3:** 243-252

Facchin F, Barbara G, Saita E, Mosconi P, Roberto A, Fedele L, Vercellini P (2015) Impact of endometriosis on quality of life and mental health: pelvic pain makes the difference. *J Psychosom Obstet Gynaecol* **36:** 135-141

Fang R, Elias DA, Monroe ME, Shen Y, McIntosh M, Wang P, Goddard CD, Callister SJ, Moore RJ, Gorby YA, Adkins JN, Fredrickson JK, Lipton MS, Smith RD (2006) Differential label-free quantitative proteomic analysis of Shewanella oneidensis cultured under aerobic and suboxic conditions by accurate mass and time tag approach. *Mol Cell Proteomics* **5:** 714-725

Farquhar C (2007) Endometriosis. *BMJ* **334:** 249-253

Farrah T, Deutsch EW, Omenn GS, Campbell DS, Sun Z, Bletz JA, Mallick P, Katz JE, Malmström J, Ossola R, Watts JD, Lin B, Zhang H, Moritz RL, Aebersold R (2011) A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol Cell Proteomics* **10:** M110.006353

Faserl K, Golderer G, Kremser L, Lindner H, Sarg B, Wildt L, Seeber B (2011) Polymorphism in vitamin D-binding protein as a genetic risk factor in the pathogenesis of endometriosis. *J Clin Endocrinol Metab* **96:** E233-241

Fassbender A, Burney RO, O DF, D'Hooghe T, Giudice L (2015) Update on Biomarkers for the Detection of Endometriosis. *Biomed Res Int* **2015:** 130854

Fassbender A, Simsa P, Kyama CM, Waelkens E, Mihalyi A, Meuleman C, Gevaert O, Van de Plas R, de Moor B, D'Hooghe TM (2010) TRIzol treatment of secretory phase endometrium allows combined proteomic and mRNA microarray analysis of the same sample in women with and without endometriosis. *Reprod Biol Endocrinol* **8:** 123

Fassbender A, Verbeeck N, Börnigen D, Kyama CM, Bokor A, Vodolazkaia A, Peeraer K, Tomassetti C, Meuleman C, Gevaert O, Van de Plas R, Ojeda F, De Moor B, Moreau Y, Waelkens E, D'Hooghe TM (2012a) Combined mRNA microarray and proteomic analysis of eutopic endometrium of women with and without endometriosis. *Hum Reprod* **27:** 2020-2029

Fassbender A, Waelkens E, Verbeeck N, Kyama CM, Bokor A, Vodolazkaia A, Van de Plas R, Meuleman C, Peeraer K, Tomassetti C, Gevaert O, Ojeda F, De Moor B, D'Hooghe T (2012b) Proteomics analysis of plasma for early diagnosis of endometriosis. *Obstet Gynecol* **119:** 276-285

Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246:** 64-71

Ferguson RE, Hochstrasser DF, Banks RE (2007) Impact of preanalytical variables on the analysis of biological fluids in proteomic studies. *Proteomics Clin Appl* **1:** 739-746

Fernández-Costa C, Calamia V, Fernández-Puente P, Capelo-Martínez JL, Ruiz-Romero C, Blanco FJ (2012) Sequential depletion of human serum for the search of osteoarthritis biomarkers. *Proteome Sci* **10:** 55

Fernández-Pérez L, Santana-Farré R, de Mirecki-Garrido M, García I, Guerra B, Mateo-Díaz C, Iglesias-Gato D, Díaz-Chico JC, Flores-Morales A, Díaz M (2014) Lipid Profiling and Transcriptomic Analysis Reveals a Functional Interplay between Estradiol and Growth Hormone in Liver. *PLoS One* **9:** e96305

Fonslow BR, Stein BD, Webb KJ, Xu T, Choi J, Park SK, Yates JR (2013) Digestion and depletion of abundant proteins improves proteomic coverage. *Nat Methods* **10:** 54-56

Fortin T, Salvador A, Charrier JP, Lenz C, Bettsworth F, Lacoux X, Choquet-Kastylevsky G, Lemoine J (2009) Multiple reaction monitoring cubed for protein quantification at the low nanogram/milliliter level in nondepleted human serum. *Anal Chem* **81:** 9343-9352

Fowler PA, Tattum J, Bhattacharya S, Klonisch T, Hombach-Klonisch S, Gazvani R, Lea RG, Miller I, Simpson WG, Cash P (2007) An investigation of the effects of endometriosis on the proteome of human eutopic endometrium: a heterogeneous tissue with a complex disease. *Proteomics* **7:** 130-142

Frank A, Pevzner P (2005) PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem* **77:** 964-973

Frantzi M, Bhat A, Latosinska A (2014) Clinical proteomic biomarkers: relevant issues on study design & technical considerations in biomarker development. *Clin Transl Med* **3:** 7

Fraser CG, Petersen PH (1999) Analytical performance characteristics should be judged against objective quality specifications. *Clin Chem* **45:** 321-323

Fredens J, Færgeman NJ (2012) Quantitative proteomics by amino acid labeling identifies novel NHR-49 regulated proteins in C. elegans. *Worm* **1:** 66-71

Friedl F, Riedl D, Fessler S, Wildt L, Walter M, Richter R, Schüßler G, Böttcher B (2015) Impact of endometriosis on quality of life, anxiety, and depression: an Austrian perspective. *Arch Gynecol Obstet* **292:** 1393-1399

Fuhrer T, Zamboni N (2015) High-throughput discovery metabolomics. *Curr Opin Biotechnol* **31:** 73-78

Fukunaga M (2000) Smooth muscle metaplasia in ovarian endometriosis. *Histopathology* **36:** 348-352

Füzéry AK, Levin J, Chan MM, Chan DW (2013) Translation of proteomic biomarkers into FDA approved cancer diagnostics: issues and challenges. *Clin Proteomics* **10:** 13

Gabbiani G (2003) The myofibroblast in wound healing and fibrocontractive diseases. *J Pathol* **200:** 500-503

Gallien S, Duriez E, Domon B (2011) Selected reaction monitoring applied to proteomics. *J Mass Spectrom* **46:** 298-312

Geiger T, Cox J, Ostasiewicz P, Wisniewski JR, Mann M (2010) Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat Methods* **7:** 383-385

Geiger T, Wehner A, Schaab C, Cox J, Mann M (2012) Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol Cell Proteomics* **11:** M111.014050

Gershon PD (2014) Cleaved and missed sites for trypsin, lys-C, and lys-N can be predicted with high confidence on the basis of sequence context. *J Proteome Res* **13:** 702-709

Gillet LC, Navarro P, Tate S, Röst H, Selevsek N, Reiter L, Bonner R, Aebersold R (2012) Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics* **11:** O111.016717

Giudice LC (2010) Clinical practice. Endometriosis. *N Engl J Med* **362:** 2389-2398

Goh WW, Lee YH, Chung M, Wong L (2012) How advancement in biological network analysis methods empowers proteomics. *Proteomics* **12:** 550-563

Gouw JW, Krijgsveld J, Heck AJ (2010) Quantitative proteomics by metabolic labeling of model organisms. *Mol Cell Proteomics* **9:** 11-24

Gregori J, Villarreal L, Sánchez A, Baselga J, Villanueva J (2013) An effect size filter improves the reproducibility in spectral counting-based comparative proteomics. *J Proteomics* **95:** 55-65

Griffin NM, Yu J, Long F, Oh P, Shore S, Li Y, Koziol JA, Schnitzer JE (2010) Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat Biotechnol* **28:** 83-89

Grossmann J, Roschitzki B, Panse C, Fortes C, Barkow-Oesterreicher S, Rutishauser D, Schlapbach R (2010) Implementation and evaluation of relative and absolute quantification in shotgun proteomics with label-free methods. *J Proteomics* **73:** 1740-1746

Güell M, Yus E, Lluch-Senar M, Serrano L (2011) Bacterial transcriptomics: what is beyond the RNA horiz-ome? *Nat Rev Microbiol* **9:** 658-669

Haas D, Wurm P, Shamiyeh A, Shebl O, Chvatal R, Oppelt P (2013) Efficacy of the revised Enzian classification: a retrospective analysis. Does the revised Enzian classification solve the problem of duplicate classification in rASRM and Enzian? *Arch Gynecol Obstet* **287:** 941-945

Hamdan M, Omar SZ, Dunselman G, Cheong Y (2015) Influence of endometriosis on assisted reproductive technology outcomes: a systematic review and meta-analysis. *Obstet Gynecol* **125:** 79-88

Hanisch FG (2012) O-glycoproteomics: site-specific O-glycoprotein analysis by CID/ETD electrospray

ionization tandem mass spectrometry and top-down glycoprotein sequencing by in-source decay MALDI mass spectrometry. *Methods Mol Biol* **842:** 179-189

Hanke S, Besir H, Oesterhelt D, Mann M (2008) Absolute SILAC for accurate quantitation of proteins in complex mixtures down to the attomole level. *J Proteome Res* **7:** 1118-1130

Harizi H, Corcuff JB, Gualde N (2008) Arachidonic-acid-derived eicosanoids: roles in biology and immunopathology. *Trends Mol Med* **14:** 461-469

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C et al (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22:** 1760-1774

Hebert AS, Richards AL, Bailey DJ, Ulbrich A, Coughlin EE, Westphall MS, Coon JJ (2014) The one hour yeast proteome. *Mol Cell Proteomics* **13:** 339-347

Hillenkamp F, Karas M (1990) Mass spectrometry of peptides and proteins by matrix-assisted ultraviolet laser desorption/ionization. *Methods Enzymol* **193:** 280-295

Holzinger ER, Ritchie MD (2012) Integrating heterogeneous high-throughput data for meta-dimensional pharmacogenomics and disease-related studies. *Pharmacogenomics* **13:** 213-222

Hoofnagle AN, Wener MH (2009) The fundamental flaws of immunoassays and potential solutions using tandem mass spectrometry. *J Immunol Methods* **347:** 3-11

Hoopmann MR, Moritz RL (2013) Current algorithmic solutions for peptide-based proteomics data generation and identification. *Curr Opin Biotechnol* **24:** 31-38

Horlacher O, Lisacek F, Müller M (2015) Mining Large Scale Tandem Mass Spectrometry Data for Protein Modifications Using Spectral Libraries. *J Proteome Res*

Hsu JL, Huang SY, Chow NH, Chen SH (2003) Stable-isotope dimethyl labeling for quantitative proteomics. *Anal Chem* **75:** 6843-6852

Hughes T, Deininger M, Hochhaus A, Branford S, Radich J, Kaeda J, Baccarani M, Cortes J, Cross NC, Druker BJ, Gabert J, Grimwade D, Hehlmann R, Kamel-Reid S, Lipton JH, Longtine J, Martinelli G, Saglio G, Soverini S, Stock W et al (2006) Monitoring CML patients responding to treatment with tyrosine kinase inhibitors: review and recommendations for harmonizing current methodology for detecting BCR-ABL transcripts and kinase domain mutations and for expressing results. *Blood* **108:** 28-37

Hutchens TW, Yip TT (1993) New Desorption Strategies for the Mass-Spectrometric Analysis of Macromolecules. *Rapid Communications in Mass Spectrometry* **7:** 576-580

Hwang JH, Lee KS, Joo JK, Wang T, Son JB, Park JH, Hwang DY, Choi MH, Lee HG (2014) Identification of biomarkers for endometriosis in plasma from patients with endometriosis using a proteomics approach. *Mol Med Rep* **10:** 725-730

Hwang JH, Wang T, Lee KS, Joo JK, Lee HG (2013) Vitamin D binding protein plays an important role in the progression of endometriosis. *Int J Mol Med* **32:** 1394-1400

Hyung SW, Kim MS, Mun DG, Lee H, Lee SW (2011) The effect and potential of using a temperature controlled separation column with ultra-high pressure microcapillary liquid chromatography/tandem mass spectrometry on proteomic analysis. *Analyst* **136:** 2100-2105

Hüttenhain R, Malmström J, Picotti P, Aebersold R (2009) Perspectives of targeted mass spectrometry for protein biomarker verification. *Curr Opin Chem Biol* **13:** 518-525

Jaffe JD, Keshishian H, Chang B, Addona TA, Gillette MA, Carr SA (2008) Accurate inclusion mass screening: a bridge from unbiased discovery to targeted assay development for biomarker verification. *Mol Cell Proteomics* **7:** 1952-1962

Jeong K, Kim S, Pevzner PA (2013) UniNovo: a universal tool for de novo peptide sequencing. *Bioinformatics* **29:** 1953-1962

Jeudy J, Salvador A, Simon R, Jaffuel A, Fonbonne C, Léonard JF, Gautier JC, Pasquier O, Lemoine J (2014) Overcoming biofluid protein complexity during targeted mass spectrometry detection and quantification of protein biomarkers by MRM cubed (MRM3). *Anal Bioanal Chem* **406:** 1193-1200

Johnstone IM, Titterington DM (2009) Statistical challenges of high-dimensional data. *Philos Trans A Math Phys Eng Sci* **367:** 4237-4253

Karp NA, Huber W, Sadowski PG, Charles PD, Hester SV, Lilley KS (2010) Addressing accuracy and precision issues in iTRAQ quantitation. *Mol Cell Proteomics* **9:** 1885-1897

Kasvandik S, Samuel K, Peters M, Eimre M, Peet N, Roost AM, Padrik L, Paju K, Peil L, Salumets A (2015) Deep Quantitative Proteomics Reveals Extensive Metabolic Reprogramming and Cancer-Like Changes of Ectopic Endometriotic Stromal Cells. *J Proteome Res*

Keich U, Kertesz-Farkas A, Noble WS (2015) Improved False Discovery Rate Estimation Procedure for Shotgun Proteomics. *J Proteome Res* **14:** 3148-3161

Khan Z, Bloom JS, Garcia BA, Singh M, Kruglyak L (2009) Protein quantification across hundreds of experimental conditions. *Proc Natl Acad Sci U S A* **106:** 15544-15548

Kim D, Shin H, Song YS, Kim JH (2012) Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *J Biomed Inform* **45:** 1191-1198

Kim MR, Park DW, Lee JH, Choi DS, Hwang KJ, Ryu HS, Min CK (2005) Progesterone-dependent release of transforming growth factor-beta1 from epithelial cells enhances the endometrial decidualization by turning on the Smad signalling in stromal cells. *Mol Hum Reprod* **11:** 801-808

Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, Thomas JK, Muthusamy B, Leal-Rojas P, Kumar P, Sahasrabuddhe NA, Balakrishnan L, Advani J, George B, Renuse S, Selvan LD et al (2014) A draft map of the human proteome. *Nature* **509:** 575-581

Kiyonami R, Schoen A, Prakash A, Peterman S, Zabrouskov V, Picotti P, Aebersold R, Huhmer A, Domon B (2011) Increased selectivity, analytical precision, and throughput in targeted proteomics. *Mol Cell Proteomics* **10:** M110.002931

Klaassen T, Szwandt S, Kapron JT, Roemer A (2009) Validated quantitation method for a peptide in rat serum using liquid chromatography/high-field asymmetric waveform ion mobility spectrometry. *Rapid Commun Mass Spectrom* **23:** 2301-2306

Kohlbacher O, Reinert K, Gröpl C, Lange E, Pfeifer N, Schulz-Trieglaff O, Sturm M (2007) TOPP--the OpenMS proteomics pipeline. *Bioinformatics* **23:** e191-197

Koni AC, Scott RA, Wang G, Bailey ME, Peplies J, Bammann K, Pitsiladis YP, Consortium I (2011) DNA yield and quality of saliva samples and suitability for large-scale epidemiological studies in children. *Int J Obes (Lond)* **35 Suppl 1:** S113-118

Koninckx PR, Kennedy SH, Barlow DH (1999) Pathogenesis of endometriosis: the role of peritoneal fluid. *Gynecol Obstet Invest* **47 Suppl 1:** 23-33

Koninckx PR, Martin DC (1992) Deep endometriosis: a consequence of infiltration or retraction or possibly adenomyosis externa? *Fertil Steril* **58:** 924-928

Krastins B, Prakash A, Sarracino DA, Nedelkov D, Niederkofler EE, Kiernan UA, Nelson R, Vogelsang MS, Vadali G, Garces A, Sutton JN, Peterman S, Byram G, Darbouret B, Pérusse JR, Seidah NG, Coulombe B, Gobom J, Portelius E, Pannee J et al (2013) Rapid development of sensitive, high-throughput, quantitative and highly selective mass spectrometric targeted immunoassays for clinically important proteins in human plasma and serum. *Clin Biochem* **46:** 399-410

Krumsiek J, Mittelstrass K, Do KT, Stückler F, Ried J, Adamski J, Peters A, Illig T, Kronenberg F, Friedrich N, Nauck M, Pietzner M, Mook-Kanamori DO, Suhre K, Gieger C, Grallert H, Theis FJ, Kastenmüller G (2015) Gender-specific pathway differences in the human serum metabolome. *Metabolomics* **11:** 1815-1833

Krüger T, Lehmann T, Rhode H (2013) Effect of quality characteristics of single sample preparation steps in the precision and coverage of proteomic studies--a review. *Anal Chim Acta* **776:** 1-10

Kyama C, T'Jampens D, Mihalyi A, Simsa P, Debrock S, Waelkens E, Landuyt B, Meuleman C, Fulop V, Mwenda J, D'Hooghe T (2006) ProteinChip technology is a useful method in the pathogenesis and diagnosis of endometriosis: a preliminary study. *Fertil Steril* **86:** 203-209

Käll L, Storey JD, MacCoss MJ, Noble WS (2008) Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J Proteome Res* **7:** 29-34

Köcher T, Pichler P, Swart R, Mechtler K (2011a) Quality control in LC-MS/MS. *Proteomics* **11:** 1026-1030

Köcher T, Swart R, Mechtler K (2011b) Ultra-high-pressure RPLC hyphenated to an LTQ-Orbitrap Velos reveals a linear relation between peak capacity and number of identified peptides. *Anal Chem* **83:** 2699-2704

Lam H, Deutsch EW, Eddes JS, Eng JK, King N, Stein SE, Aebersold R (2007) Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7:** 655-667

Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* **3:** 1724-1735

Leptos KC, Sarracino DA, Jaffe JD, Krastins B, Church GM (2006) MapQuant: open-source software for large-scale protein quantification. *Proteomics* **6:** 1770-1782

Li X, Nokkala E, Yan W, Streng T, Saarinen N, Wärri A, Huhtaniemi I, Santti R, Mäkelä S, Poutanen M (2001) Altered structure and function of reproductive organs in transgenic male mice overexpressing human aromatase. *Endocrinology* **142:** 2435-2442

Li X, Strauss L, Kaatrasalo A, Mayerhofer A, Huhtaniemi I, Santti R, Mäkelä S, Poutanen M (2006) Transgenic mice expressing p450 aromatase as a model for male infertility associated with chronic inflammation in the testis. *Endocrinology* **147:** 1271-1277

Li X, Strauss L, Mäkelä S, Streng T, Huhtaniemi I, Santti R, Poutanen M (2004) Multiple structural and functional abnormalities in the p450 aromatase expressing transgenic male mice are ameliorated by a p450 aromatase inhibitor. *Am J Pathol* **164:** 1039-1048

Li X, Wärri A, Mäkelä S, Ahonen T, Streng T, Santti R, Poutanen M (2002) Mammary gland development in transgenic male mice expressing human P450 aromatase. *Endocrinology* **143:** 4074-4083

Li Z, Adams RM, Chourey K, Hurst GB, Hettich RL, Pan C (2012) Systematic comparison of label-free, metabolic labeling, and isobaric chemical labeling

for quantitative proteomics on LTQ Orbitrap Velos. *J Proteome Res* **11:** 1582-1590

Licklider LJ, Thoreen CC, Peng J, Gygi SP (2002) Automation of nanoscale microcapillary liquid chromatography-tandem mass spectrometry with a vented column. *Anal Chem* **74:** 3076-3083

Liu B, Qiu FH, Voss C, Xu Y, Zhao MZ, Wu YX, Nie J, Wang ZL (2011) Evaluation of three high abundance protein depletion kits for umbilical cord serum proteomics. *Proteome Sci* **9:** 24

Liu H, Lang J, Zhou Q, Shan D, Li Q (2007) Detection of endometriosis with the use of plasma protein profiling by surface-enhanced laser desorption/ ionization time-of-flight mass spectrometry. *Fertil Steril* **87:** 988-990

Liu H, Sadygov RG, Yates JR (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **76:** 4193-4201

Liu Y, Hüttenhain R, Surinova S, Gillet LC, Mouritsen J, Brunner R, Navarro P, Aebersold R (2013) Quantitative measurements of N-linked glycoproteins in human plasma by SWATH-MS. *Proteomics* **13:** 1247-1256

Lo Turco EG, Cordeiro FB, Lopes PH, Gozzo FC, Pilau EJ, Soler TB, da Silva BF, Del Giudice PT, Bertolla RP, Fraietta R, Cedenho AP (2013) Proteomic analysis of follicular fluid from women with and without endometriosis: new therapeutic targets and biomarkers. *Mol Reprod Dev* **80:** 441-450

Long X, Jiang P, Zhou L, Zhang W (2013) Evaluation of novel serum biomarkers and the proteomic differences of endometriosis and adenomyosis using MALDI-TOF-MS. *Arch Gynecol Obstet* **288:** 201-205

Longoria TC, Ueland FR, Zhang Z, Chan DW, Smith A, Fung ET, Munroe DG, Bristow RE (2014) Clinical performance of a multivariate index assay for detecting early-stage ovarian cancer. *Am J Obstet Gynecol* **210:** 78.e71-79

Low TY, van Heesch S, van den Toorn H, Giansanti P, Cristobal A, Toonen P, Schafer S, Hübner N, van Breukelen B, Mohammed S, Cuppen E, Heck AJ, Guryev V (2013) Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. *Cell Rep* **5:** 1469-1478

Lu Y, Liu X, Jiao Y, Xiong X, Wang E, Wang X, Zhang Z, Zhang H, Pan L, Guan Y, Cai D, Ning G, Li X (2014) Periostin promotes liver steatosis and hypertriglyceridemia through downregulation of PPARα. *J Clin Invest* **124:** 3501-3513

Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* **17:** 2337-2342

Ma ZQ, Polzin KO, Dasari S, Chambers MC, Schilling B, Gibson BW, Tran BQ, Vega-Montoto L, Liebler DC, Tabb DL (2012) QuaMeter: multivendor performance metrics for LC-MS/MS proteomics instrumentation. *Anal Chem* **84:** 5845-5850

MacCoss MJ, Wu CC, Liu H, Sadygov R, Yates JR (2003) A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. *Anal Chem* **75:** 6912-6921

Mackowiak SD, Zauber H, Bielow C, Thiel D, Kutz K, Calviello L, Mastrobuoni G, Rajewsky N, Kempa S, Selbach M, Obermayer B (2015) Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol* **16:** 179

Mahn A, Ismail M (2011) Depletion of highly abundant proteins in blood plasma by ammonium sulfate precipitation for 2D-PAGE analysis. *J Chromatogr B Analyt Technol Biomed Life Sci* **879:** 3645-3648

Makarov A, Denisov E, Kholomeev A, Balschun W, Lange O, Strupat K, Horning S (2006) Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal Chem* **78:** 2113-2120

Malaud E, Piquer D, Merle D, Molina L, Guerrier L, Boschetti E, Saussine M, Marty-Ané C, Albat B, Fareh J (2012) Carotid atherosclerotic plaques: proteomics study after a low-abundance protein enrichment step. *Electrophoresis* **33:** 470-482

Mann B, Madera M, Sheng Q, Tang H, Mechref Y, Novotny MV (2008) ProteinQuant Suite: a bundle of automated software tools for label-free quantitative proteomics. *Rapid Commun Mass Spectrom* **22:** 3823-3834

Marcelo KL, Lin F, Rajapakshe K, Dean A, Gonzales N, Coarfa C, Means AR, Goldie LC, York B (2015) Deciphering hepatocellular responses to metabolic and oncogenic stress. *J Biol Methods* **2**

Marianowski P, Szymusik I, Malejczyk J, Hibner M, Wielgos M (2013) Proteomic analysis of eutopic and ectopic endometriotic tissues based on isobaric peptide tags for relative and absolute quantification (iTRAQ) method. *Neuro Endocrinol Lett* **34:** 717-721

Martínez-Bartolomé S, Deutsch EW, Binz PA, Jones AR, Eisenacher M, Mayer G, Campos A, Canals F, Bech-Serra JJ, Carrascal M, Gay M, Paradela A, Navajas R, Marcilla M, Hernáez ML, Gutiérrez-Blázquez MD, Velarde LF, Aloria K, Beaskoetxea J, Medina-Aunon JA et al (2013) Guidelines for reporting quantitative mass spectrometry based experiments in proteomics. *J Proteomics* **95:** 84-88

May KE, Villar J, Kirtley S, Kennedy SH, Becker CM (2011) Endometrial alterations in endometriosis: a systematic review of putative biomarkers. *Hum Reprod Update* **17:** 637-653

McAlister GC, Berggren WT, Griep-Raming J, Horning S, Makarov A, Phanstiel D, Stafford G, Swaney DL, Syka JE, Zabrouskov V, Coon JJ (2008) A proteomics grade electron transfer dissociation-enabled hybrid linear

ion trap-orbitrap mass spectrometer. *J Proteome Res* **7:** 3127-3136

McAlister GC, Phanstiel D, Good DM, Berggren WT, Coon JJ (2007) Implementation of electron-transfer dissociation on a hybrid linear ion trap-orbitrap mass spectrometer. *Anal Chem* **79:** 3525-3534

McNulty DE, Annan RS (2008) Hydrophilic interaction chromatography reduces the complexity of the phosphoproteome and improves global phosphopeptide isolation and detection. *Mol Cell Proteomics* **7:** 971-980

Medvedev A, Kopylov A, Buneeva O, Zgoda V, Archakov A (2012) Affinity-based proteomic profiling: problems and achievements. *Proteomics* **12:** 621-637

Michalski A, Cox J, Mann M (2011a) 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* **10:** 1785-1793

Michalski A, Damoc E, Hauschild JP, Lange O, Wieghaus A, Makarov A, Nagaraj N, Cox J, Mann M, Horning S (2011b) Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol Cell Proteomics* **10:** M111.011015

Michalski A, Damoc E, Lange O, Denisov E, Nolting D, Müller M, Viner R, Schwartz J, Remes P, Belford M, Dunyach JJ, Cox J, Horning S, Mann M, Makarov A (2012) Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Mol Cell Proteomics* **11:** O111.013698

Milan E, Lazzari C, Anand S, Floriani I, Torri V, Sorlini C, Gregorc V, Bachi A (2012) SAA1 is over-expressed in plasma of non small cell lung cancer patients with poor outcome after treatment with epidermal growth factor receptor tyrosine-kinase inhibitors. *J Proteomics* **76 Spec No.:** 91-101

Millioni R, Tolin S, Puricelli L, Sbrignadello S, Fadini GP, Tessari P, Arrigoni G (2011) High abundance proteins depletion vs low abundance proteins enrichment: comparison of methods to reduce the plasma proteome complexity. *PLoS One* **6:** e19603

Milner E, Barnea E, Beer I, Admon A (2006) The turnover kinetics of major histocompatibility complex peptides of human cancer cells. *Mol Cell Proteomics* **5:** 357-365

Mittelstrass K, Ried JS, Yu Z, Krumsiek J, Gieger C, Prehn C, Roemisch-Margl W, Polonikov A, Peters A, Theis FJ, Meitinger T, Kronenberg F, Weidinger S, Wichmann HE, Suhre K, Wang-Sattler R, Adamski J, Illig T (2011) Discovery of sexual dimorphisms in metabolic and genetic biomarkers. *PLoS Genet* **7:** e1002215

Molina H, Yang Y, Ruch T, Kim JW, Mortensen P, Otto T, Nalli A, Tang QQ, Lane MD, Chaerkady R, Pandey A (2009) Temporal profiling of the adipocyte proteome during differentiation using a five-plex SILAC based strategy. *J Proteome Res* **8:** 48-58

Monroe ME, Tolić N, Jaitly N, Shaw JL, Adkins JN, Smith RD (2007) VIPER: an advanced software package to support high-throughput LC-MS peptide identification. *Bioinformatics* **23:** 2021-2023

Motoyama A, Venable JD, Ruse CI, Yates JR (2006) Automated ultra-high-pressure multidimensional protein identification technology (UHP-MudPIT) for improved peptide identification of proteomic samples. *Anal Chem* **78:** 5109-5118

Mueller LN, Rinner O, Schmidt A, Letarte S, Bodenmiller B, Brusniak MY, Vitek O, Aebersold R, Müller M (2007) SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* **7:** 3470-3480

Muller DN, Schmidt C, Barbosa-Sicard E, Wellner M, Gross V, Hercule H, Markovic M, Honeck H, Luft FC, Schunck WH (2007) Mouse Cyp4a isoforms: enzymatic properties, gender- and strain-specific expression, and role in renal 20-hydroxyeicosatetraenoic acid formation. *Biochem J* **403:** 109-118

Murphy JP, Stepanova E, Everley RA, Paulo JA, Gygi SP (2015) Comprehensive Temporal Protein Dynamics during the Diauxic Shift in Saccharomyces cerevisiae. *Mol Cell Proteomics* **14:** 2454-2465

Nagaraj N, Kulak NA, Cox J, Neuhauser N, Mayr K, Hoerning O, Vorm O, Mann M (2012) 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* **11:** M111.013722

Nagaraj N, Mann M (2011) Quantitative analysis of the intra- and inter-individual variability of the normal urinary proteome. *J Proteome Res* **10:** 637-645

Nahnsen S, Bielow C, Reinert K, Kohlbacher O (2013) Tools for label-free peptide quantification. *Mol Cell Proteomics* **12:** 549-556

Nemoto Y, Toda K, Ono M, Fujikawa-Adachi K, Saibara T, Onishi S, Enzan H, Okada T, Shizuta Y (2000) Altered expression of fatty acid-metabolizing enzymes in aromatase-deficient mice. *J Clin Invest* **105:** 1819-1825

Nesvizhskii AI, Aebersold R (2005) Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* **4:** 1419-1440

Nesvizhskii AI, Roos FF, Grossmann J, Vogelzang M, Eddes JS, Gruissem W, Baginsky S, Aebersold R (2006) Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol Cell Proteomics* **5:** 652-670

Nielsen ML, Savitski MM, Zubarev RA (2006) Extent of modifications in human proteome samples and their effect on dynamic range of analysis in shotgun proteomics. *Mol Cell Proteomics* **5:** 2384-2391

Ning K, Fermin D, Nesvizhskii AI (2012) Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data. *J Proteome Res* **11:** 2261-2271

Nnoaham KE, Hummelshoj L, Webster P, d'Hooghe T, de Cicco Nardone F, de Cicco Nardone C, Jenkinson C, Kennedy SH, Zondervan KT, consortium WERFGSoWsH (2011) Impact of endometriosis on quality of life and work productivity: a multicenter study across ten countries. *Fertil Steril* **96:** 366-373. e368

Obuchowski NA, Lieber ML, Wians FH (2004) ROC curves in clinical chemistry: uses, misuses, and possible solutions. *Clin Chem* **50:** 1118-1125

Oda Y, Huang K, Cross FR, Cowburn D, Chait BT (1999) Accurate quantitation of protein expression and site-specific phosphorylation. *Proc Natl Acad Sci U S A* **96:** 6591-6596

Old WM, Meyer-Arendt K, Aveline-Wolf L, Pierce KG, Mendoza A, Sevinsky JR, Resing KA, Ahn NG (2005) Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol Cell Proteomics* **4:** 1487-1502

Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **1:** 376-386

Ong SE, Mann M (2005) Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol* **1:** 252-262

Oppermann FS, Klammer M, Bobe C, Cox J, Schaab C, Tebbe A, Daub H (2013) Comparison of SILAC and mTRAQ quantification for phosphoproteomics on a quadrupole orbitrap mass spectrometer. *J Proteome Res* **12:** 4089-4100

Palmer BF, Clegg DJ (2015) The sexual dimorphism of obesity. *Mol Cell Endocrinol* **402:** 113-119

Park SK, Venable JD, Xu T, Yates JR (2008) A quantitative analysis software tool for mass spectrometry-based proteomics. *Nat Methods* **5:** 319-322

Parker CE, Domanski D, Percy AJ, Chambers AG, Camenzind AG, Smith DS, Borchers CH (2014) Mass spectrometry in high-throughput clinical biomarker assays: multiple reaction monitoring. *Top Curr Chem* **336:** 117-137

Patel BB, Barrero CA, Braverman A, Kim PD, Jones KA, Chen DE, Bowler RP, Merali S, Kelsen SG, Yeung AT (2012) Assessment of two immunodepletion methods: off-target effects and variations in immunodepletion efficiency may confound plasma proteomics. *J Proteome Res* **11:** 5947-5958

Paulovich AG, Billheimer D, Ham AJ, Vega-Montoto L, Rudnick PA, Tabb DL, Wang P, Blackman RK, Bunk DM, Cardasis HL, Clauser KR, Kinsinger CR, Schilling B, Tegeler TJ, Variyath AM, Wang M, Whiteaker JR, Zimmerman LJ, Fenyo D, Carr SA et al (2010) Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. *Mol Cell Proteomics* **9:** 242-254

Pelander A, Decker P, Baessmann C, Ojanperä I (2011) Evaluation of a high resolving power time-of-flight mass spectrometer for drug analysis in terms of resolving power and acquisition rate. *J Am Soc Mass Spectrom* **22:** 379-385

Percy AJ, Chambers AG, Yang J, Hardie DB, Borchers CH (2014) Advances in multiplexed MRM-based protein biomarker quantitation toward clinical utility. *Biochim Biophys Acta* **1844:** 917-926

Percy AJ, Chambers AG, Yang J, Jackson AM, Domanski D, Burkhart J, Sickmann A, Borchers CH (2013a) Method and platform standardization in MRM-based quantitative plasma proteomics. *J Proteomics* **95:** 66-76

Percy AJ, Parker CE, Borchers CH (2013b) Pre-analytical and analytical variability in absolute quantitative MRM-based plasma proteomic studies. *Bioanalysis* **5:** 2837-2856

Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20:** 3551-3567

Peterson AC, Russell JD, Bailey DJ, Westphall MS, Coon JJ (2012) Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Mol Cell Proteomics* **11:** 1475-1488

Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, Liotta LA (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359:** 572-577

Petricoin EF, Belluco C, Araujo RP, Liotta LA (2006) The blood peptidome: a higher dimension of information content for cancer biomarker discovery. *Nat Rev Cancer* **6:** 961-967

Picotti P, Lam H, Campbell D, Deutsch EW, Mirzaei H, Ranish J, Domon B, Aebersold R (2008) A database of mass spectrometric assays for the yeast proteome. *Nat Methods* **5:** 913-914

Piehowski PD, Petyuk VA, Orton DJ, Xie F, Moore RJ, Ramirez-Restrepo M, Engel A, Lieberman AP, Albin RL, Camp DG, Smith RD, Myers AJ (2013) Sources of technical variability in quantitative LC-MS proteomics: human brain tissue sample analysis. *J Proteome Res* **12:** 2128-2137

Piltonen T, Puurunen J, Hedberg P, Ruokonen A, Mutt SJ, Herzig KH, Nissinen A, Morin-Papunen L, Tapanainen JS (2012) Oral, transdermal and vaginal combined contraceptives induce an increase in markers of chronic inflammation and impair insulin sensitivity in young healthy normal-weight women: a randomized study. *Hum Reprod* **27:** 3046-3056

Pluskal T, Castillo S, Villar-Briones A, Oresic M (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11:** 395

Polanski M, Anderson NL (2007) A list of candidate cancer biomarkers for targeted proteomics. *Biomark Insights* **1:** 1-48

Prakash A, Rezai T, Krastins B, Sarracino D, Athanas M, Russo P, Zhang H, Tian Y, Li Y, Kulasingam V, Drabovich A, Smith CR, Batruch I, Oran PE, Fredolini C, Luchini A, Liotta L, Petricoin E, Diamandis EP, Chan DW et al (2012) Interlaboratory reproducibility of selective reaction monitoring assays using multiple upfront analyte enrichment strategies. *J Proteome Res* **11:** 3986-3995

Pupo-Nogueira A, de Oliveira RM, Petta CA, Podgaec S, Dias JA, Abrao MS (2007) Vascular endothelial growth factor concentrations in the serum and peritoneal fluid of women with endometriosis. *Int J Gynaecol Obstet* **99:** 33-37

Qian WJ, Liu T, Petyuk VA, Gritsenko MA, Petritis BO, Polpitiya AD, Kaushal A, Xiao W, Finnerty CC, Jeschke MG, Jaitly N, Monroe ME, Moore RJ, Moldawer LL, Davis RW, Tompkins RG, Herndon DN, Camp DG, Smith RD, Program IatHRtILSCR (2009) Large-scale multiplexed quantitative discovery proteomics enabled by the use of an (18)O-labeled "universal" reference sample. *J Proteome Res* **8:** 290-299

Radulovic D, Jelveh S, Ryu S, Hamilton TG, Foss E, Mao Y, Emili A (2004) Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. *Mol Cell Proteomics* **3:** 984-997

Rai P, Kota V, Deendayal M, Shivaji S (2010) Differential proteome profiling of eutopic endometrium from women with endometriosis to understand etiology of endometriosis. *J Proteome Res* **9:** 4407-4419

Ralton LD, Murray GI (2011) The use of formalin fixed wax embedded tissue for proteomic analysis. *J Clin Pathol* **64:** 297-302

Ransohoff DF (2004) Rules of evidence for cancer molecular-marker discovery and validation. *Nat Rev Cancer* **4:** 309-314

Ransohoff DF, Gourlay ML (2010) Sources of bias in specimens for research about molecular markers for cancer. *J Clin Oncol* **28:** 698-704

Raychaudhuri S, Plenge RM, Rossin EJ, Ng AC, Purcell SM, Sklar P, Scolnick EM, Xavier RJ, Altshuler D, Daly MJ, Consortium IS (2009) Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet* **5:** e1000534

Richards AL, Merrill AE, Coon JJ (2015) Proteome sequencing goes deep. *Curr Opin Chem Biol* **24:** 11-17

Rifai N, Gillette MA, Carr SA (2006) Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat Biotechnol* **24:** 971-983

Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D (2015) Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet* **16:** 85-97

Roepstorff P, Fohlman J (1984) Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom* **11:** 601

Rogers PA, D'Hooghe TM, Fazleabas A, Giudice LC, Montgomery GW, Petraglia F, Taylor RN (2013) Defining future directions for endometriosis research: workshop report from the 2011 World Congress of Endometriosis In Montpellier, France. *Reprod Sci* **20:** 483-499

Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, Purkayastha S, Juhasz P, Martin S, Bartlet-Jones M, He F, Jacobson A, Pappin DJ (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* **3:** 1154-1169

Rudnick PA, Clauser KR, Kilpatrick LE, Tchekhovskoi DV, Neta P, Blonder N, Billheimer DD, Blackman RK, Bunk DM, Cardasis HL, Ham AJ, Jaffe JD, Kinsinger CR, Mesri M, Neubert TA, Schilling B, Tabb DL, Tegeler TJ, Vega-Montoto L, Variyath AM et al (2010) Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. *Mol Cell Proteomics* **9:** 225-241

Russell MR, Lilley KS (2012) Pipeline to assess the greatest source of technical variance in quantitative proteomics using metabolic labelling. *J Proteomics* **77:** 441-454

Rybakova KN, Bruggeman FJ, Tomaszewska A, Moné MJ, Carlberg C, Westerhoff HV (2015) Multiplex Eukaryotic Transcription (In)activation: Timing, Bursting and Cycling of a Ratchet Clock Mechanism. *PLoS Comput Biol* **11:** e1004236

Sackett DL, Haynes RB (2002) The architecture of diagnostic research. *BMJ* **324:** 539-541

Sandin M, Ali A, Hansson K, Månsson O, Andreasson E, Resjö S, Levander F (2013) An adaptive alignment algorithm for quality-controlled label-free LC-MS. *Mol Cell Proteomics* **12:** 1407-1420

Sandin M, Chawade A, Levander F (2015) Is label-free LC-MS/MS ready for biomarker discovery? *Proteomics Clin Appl* **9:** 289-294

Sandin M, Krogh M, Hansson K, Levander F (2011) Generic workflow for quality assessment of quantitative label-free LC-MS analysis. *Proteomics* **11:** 1114-1124

Sandra K, Moshir M, D'hondt F, Tuytten R, Verleysen K, Kas K, François I, Sandra P (2009) Highly efficient peptide separations in proteomics. Part 2: bi- and multidimensional liquid-based separation techniques. *J Chromatogr B Analyt Technol Biomed Life Sci* **877:** 1019-1039

Sandra K, Moshir M, D'hondt F, Verleysen K, Kas K, Sandra P (2008) Highly efficient peptide separations in proteomics Part 1. Unidimensional high performance liquid chromatography. *J Chromatogr B Analyt Technol Biomed Life Sci* **866:** 48-63

Sarachana T, Hu VW (2013) Differential recruitment of coregulators to the RORA promoter adds another layer of complexity to gene (dys) regulation by sex hormones in autism. *Mol Autism* **4:** 39

Sarbu M, Ghiulai RM, Zamfir AD (2014) Recent developments and applications of electron transfer dissociation mass spectrometry in proteomics. *Amino Acids* **46:** 1625-1634

Savitski MM, Wilhelm M, Hahne H, Kuster B, Bantscheff M (2015) A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. *Mol Cell Proteomics* **14:** 2394-2404

Scheltema RA, Mann M (2012) SprayQc: a real-time LC-MS/MS quality monitoring system to maximize uptime using off the shelf components. *J Proteome Res* **11:** 3458-3466

Schilling B, MacLean B, Held JM, Sahu AK, Rardin MJ, Sorensen DJ, Peters T, Wolfe AJ, Hunter CL, MacCoss MJ, Gibson BW (2015) Multiplexed, Scheduled, High-Resolution Parallel Reaction Monitoring on a Full Scan QqTOF Instrument with Integrated Data-Dependent and Targeted Mass Spectrometric Workflows. *Anal Chem* **87:** 10222-10229

Schmidt A, Gehlenborg N, Bodenmiller B, Mueller LN, Campbell D, Mueller M, Aebersold R, Domon B (2008) An integrated, directed mass spectrometric approach for in-depth characterization of complex peptide mixtures. *Mol Cell Proteomics* **7:** 2138-2150

Schrag D, Kuntz KM, Garber JE, Weeks JC (2000) Life expectancy gains from cancer prevention strategies for women with breast cancer and BRCA1 or BRCA2 mutations. *JAMA* **283:** 617-624

Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M (2011) Global quantification of mammalian gene expression control. *Nature* **473:** 337-342

Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M (2013) Corrigendum: Global quantification of mammalian gene expression control. *Nature* **495:** 126-127

Schwanhäusser B, Gossen M, Dittmar G, Selbach M (2009) Global analysis of cellular protein translation by pulsed SILAC. *Proteomics* **9:** 205-209

Scigelova M, Makarov A (2006) Orbitrap mass analyzer--overview and applications in proteomics. *Proteomics* **6 Suppl 2:** 16-21

Seeber B, Sammel MD, Fan X, Gerton GL, Shaunik A, Chittams J, Barnhart KT (2010) Proteomic analysis of serum yields six candidate proteins that are differentially regulated in a subset of women with endometriosis. *Fertil Steril* **93:** 2137-2144

Selvaraju S, Rassi ZE (2012) Liquid-phase-based separation systems for depletion, prefractionation and enrichment of proteins in biological fluids and matrices for in-depth proteomics analysis--an update covering the period 2008-2011. *Electrophoresis* **33:** 74-88

Senko MW, Remes PM, Canterbury JD, Mathur R, Song Q, Eliuk SM, Mullen C, Earley L, Hardman M, Blethrow JD, Bui H, Specht A, Lange O, Denisov E, Makarov A, Horning S, Zabrouskov V (2013) Novel parallelized quadrupole/linear ion trap/Orbitrap tribrid mass spectrometer improving proteome coverage and peptide identification rates. *Anal Chem* **85:** 11710-11714

Sheldon E, Vo KC, McIntire RA, Aghajanova L, Zelenko Z, Irwin JC, Giudice LC (2011) Biobanking human endometrial tissue and blood specimens: standard operating procedure and importance to reproductive biology research and diagnostic development. *Fertil Steril* **95:** 2120-2122, 2122. e2121-2112

Shi R, Kumar C, Zougman A, Zhang Y, Podtelejnikov A, Cox J, Wiśniewski JR, Mann M (2007) Analysis of the mouse liver proteome using advanced mass spectrometry. *J Proteome Res* **6:** 2963-2972

Shi T, Zhou JY, Gritsenko MA, Hossain M, Camp DG, Smith RD, Qian WJ (2012) IgY14 and SuperMix immunoaffinity separations coupled with liquid chromatography-mass spectrometry for human plasma proteomics biomarker discovery. *Methods* **56:** 246-253

Shirtcliff EA, Granger DA, Schwartz E, Curran MJ (2001) Use of salivary biomarkers in biobehavioral research: cotton-based sample collection methods can interfere with salivary immunoassay results. *Psychoneuroendocrinology* **26:** 165-173

Shteynberg D, Nesvizhskii AI, Moritz RL, Deutsch EW (2013) Combining results of multiple search engines in proteomics. *Mol Cell Proteomics* **12:** 2383-2393

Shuford CM, Hawkridge AM, Burnett JC, Muddiman DC (2010) Utilizing spectral counting to quantitatively characterize tandem removal of abundant proteins (TRAP) in human plasma. *Anal Chem* **82:** 10179-10185

Silva JC, Denny R, Dorschel CA, Gorenstein M, Kass IJ, Li GZ, McKenna T, Nold MJ, Richardson K, Young

P, Geromanos S (2005) Quantitative proteomic analysis by accurate mass retention time pairs. *Anal Chem* **77:** 2187-2200

Sohler F, Sommer A, Wachter DL, Agaimy A, Fischer OM, Renner SP, Burghaus S, Fasching PA, Beckmann MW, Fuhrmann U, Strick R, Strissel PL (2013) Tissue remodeling and nonendometrium-like menstrual cycling are hallmarks of peritoneal endometriosis lesions. *Reprod Sci* **20:** 85-102

Solberg HE (1987) International Federation of Clinical Chemistry (IFCC), Scientific Committee, Clinical Section, Expert Panel on Theory of Reference Values, and International Committee for Standardization in Haematology (ICSH), Standing Committee on Reference Values. Approved Recommendation (1986) on the theory of reference values. Part 1. The concept of reference values. *J Clin Chem Clin Biochem* **25:** 337-342

Solberg HE, PetitClerc C (1988) International Federation of Clinical Chemistry (IFCC), Scientific Committee, Clinical Section, Expert Panel on Theory of Reference Values. Approved recommendation (1988) on the theory of reference values. Part 3. Preparation of individuals and collection of specimens for the production of reference values. *J Clin Chem Clin Biochem* **26:** 593-598

Soliman AM, Yang H, Du EX, Kelley C, Winkel C (2016) The direct and indirect costs associated with endometriosis: a systematic literature review. *Hum Reprod* **31:** 712-722

States DJ, Omenn GS, Blackwell TW, Fermin D, Eng J, Speicher DW, Hanash SM (2006) Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. *Nat Biotechnol* **24:** 333-338

Stenman UH (2013) Validating serum markers for monitoring of cancer. *Clin Chem* **59:** 4-5

Stephens AN, Hannan NJ, Rainczuk A, Meehan KL, Chen J, Nicholls PK, Rombauts LJ, Stanton PG, Robertson DM, Salamonsen LA (2010) Post-translational modifications and protein-specific isoforms in endometriosis revealed by 2D DIGE. *J Proteome Res* **9:** 2438-2449

Sturm M, Bertsch A, Gröpl C, Hildebrandt A, Hussong R, Lange E, Pfeifer N, Schulz-Trieglaff O, Zerck A, Reinert K, Kohlbacher O (2008) OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinformatics* **9:** 163

Sun A, Jiang Y, Wang X, Liu Q, Zhong F, He Q, Guan W, Li H, Sun Y, Shi L, Yu H, Yang D, Xu Y, Song Y, Tong W, Li D, Lin C, Hao Y, Geng C, Yun D et al (2010) Liverbase: a comprehensive view of human liver biology. *J Proteome Res* **9:** 50-58

Sun W, Wu S, Wang X, Zheng D, Gao Y (2005) An analysis of protein abundance suppression in data dependent liquid chromatography and tandem mass spectrometry with tryptic peptide mixtures of five known proteins. *Eur J Mass Spectrom (Chichester, Eng)* **11:** 575-580

Syka JE, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A* **101:** 9528-9533

Tabb DL, Vega-Montoto L, Rudnick PA, Variyath AM, Ham AJ, Bunk DM, Kilpatrick LE, Billheimer DD, Blackman RK, Cardasis HL, Carr SA, Clauser KR, Jaffe JD, Kowalski KA, Neubert TA, Regnier FE, Schilling B, Tegeler TJ, Wang M, Wang P et al (2010) Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J Proteome Res* **9:** 761-776

Taouatas N, Altelaar AF, Drugan MM, Helbig AO, Mohammed S, Heck AJ (2009) Strong cation exchange-based fractionation of Lys-N-generated peptides facilitates the targeted analysis of post-translational modifications. *Mol Cell Proteomics* **8:** 190-200

Taylor CF, Binz PA, Aebersold R, Affolter M, Barkovich R, Deutsch EW, Horn DM, Hühmer A, Kussmann M, Lilley K, Macht M, Mann M, Müller D, Neubert TA, Nickson J, Patterson SD, Raso R, Resing K, Seymour SL, Tsugita A et al (2008) Guidelines for reporting the use of mass spectrometry in proteomics. *Nat Biotechnol* **26:** 860-861

Taylor RM, Dance J, Taylor RJ, Prince JT (2013) Metriculator: quality assessment for mass spectrometry-based proteomics. *Bioinformatics* **29:** 2948-2949

Ten Have S, Fraser I, Markham R, Lam A, Matsumoto I (2007) Proteomic analysis of protein expression in the eutopic endometrium of women with endometriosis. *Proteomics Clin Appl* **1:** 1243-1251

Thakur SS, Geiger T, Chatterjee B, Bandilla P, Fröhlich F, Cox J, Mann M (2011) Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Mol Cell Proteomics* **10:** M110.003699

Thompson A, Schäfer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Johnstone R, Mohammed AK, Hamon C (2003) Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **75:** 1895-1904

Ting L, Cowley MJ, Hoon SL, Guilhaus M, Raftery MJ, Cavicchioli R (2009) Normalization and statistical analysis of quantitative proteomics data generated by metabolic labeling. *Mol Cell Proteomics* **8:** 2227-2242

Ting L, Rad R, Gygi SP, Haas W (2011) MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat Methods* **8:** 937-940

Tokushige N, Markham R, Crossett B, Ahn SB, Nelaturi VL, Khan A, Fraser IS (2011) Discovery

of a novel biomarker in the urine in women with endometriosis. *Fertil Steril* **95:** 46-49

Tokushige N, Markham R, Russell P, Fraser IS (2006) Nerve fibres in peritoneal endometriosis. *Hum Reprod* **21:** 3001-3007

Tsou CC, Tsai CF, Tsui YH, Sudhir PR, Wang YT, Chen YJ, Chen JY, Sung TY, Hsu WL (2010) IDEAL-Q, an automated tool for label-free quantitation analysis using an efficient peptide alignment approach and spectral data validation. *Mol Cell Proteomics* **9:** 131-144

Tu C, Li J, Young R, Page BJ, Engler F, Halfon MS, Canty JM, Qu J (2011) Combinatorial peptide ligand library treatment followed by a dual-enzyme, dual-activation approach on a nanoflow liquid chromatography/orbitrap/electron transfer dissociation system for comprehensive analysis of swine plasma proteome. *Anal Chem* **83:** 4802-4813

Tu C, Rudnick PA, Martinez MY, Cheek KL, Stein SE, Slebos RJ, Liebler DC (2010) Depletion of abundant plasma proteins and limitations of plasma proteomics. *J Proteome Res* **9:** 4982-4991

Tuerxun H, Zhang Y, Ji F, Aili A, Yang X, Ding Y (2014) Application of two-dimensional electrophoresis and mass spectrometry to screen endometriosis-related proteins. *Mol Med Rep* **10:** 95-100

Tworoger SS, Hankinson SE (2006) Use of biomarkers in epidemiologic studies: minimizing the influence of measurement error in the study design and analysis. *Cancer Causes Control* **17:** 889-899

Unlü M, Morgan M, Minden J (1997) Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis* **18:** 2071-2077

van Kaam KJ, Schouten JP, Nap AW, Dunselman GA, Groothuis PG (2008) Fibromuscular differentiation in deeply infiltrating endometriosis is a reaction of resident fibroblasts to the presence of ectopic endometrium. *Hum Reprod* **23:** 2692-2700

Van Poucke M, Vanhaesebrouck AE, Peelman LJ, Van Ham L (2012) Experimental validation of in silico predicted KCNA1, KCNA2, KCNA6 and KCNQ2 genes for association studies of peripheral nerve hyperexcitability syndrome in Jack Russell Terriers. *Neuromuscul Disord* **22:** 558-565

Vandermarliere E, Mueller M, Martens L (2013) Getting intimate with trypsin, the leading protease in proteomics. *Mass Spectrom Rev* **32:** 453-465

Varnum SM, Webb-Robertson BJ, Hessol NA, Smith RD, Zangar RC (2011) Plasma biomarkers for detecting Hodgkin's lymphoma in HIV patients. *PLoS One* **6:** e29263

Vasilj A, Gentzel M, Ueberham E, Gebhardt R, Shevchenko A (2012) Tissue proteomics by one-dimensional gel electrophoresis combined with label-free protein quantification. *J Proteome Res* **11:** 3680-3689

Vercellini P, Fedele L, Aimi G, Pietropaolo G, Consonni D, Crosignani PG (2007) Association between endometriosis stage, lesion type, patient characteristics and severity of pelvic pain symptoms: a multivariate analysis of over 1000 patients. *Hum Reprod* **22:** 266-271

Villanueva J, Carrascal M, Abian J (2014) Isotope dilution mass spectrometry for absolute quantification in proteomics: concepts and strategies. *J Proteomics* **96:** 184-199

Vogel C, Marcotte EM (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* **13:** 227-232

von Toerne C, Kahle M, Schäfer A, Ispiryan R, Blindert M, Hrabe De Angelis M, Neschen S, Ueffing M, Hauck SM (2013) Apoe, Mbl2, and Psp plasma protein levels correlate with diabetic phenotype in NZO mice--an optimized rapid workflow for SRM-based quantification. *J Proteome Res* **12:** 1331-1343

Waller KG, Lindsay P, Curtis P, Shaw RW (1993) The prevalence of endometriosis in women with infertile partners. *Eur J Obstet Gynecol Reprod Biol* **48:** 135-139

Wang L, Liu HY, Shi HH, Lang JH, Sun W (2014) Urine peptide patterns for non-invasive diagnosis of endometriosis: a preliminary prospective study. *Eur J Obstet Gynecol Reprod Biol* **177:** 23-28

Wang P, Whiteaker JR, Paulovich AG (2009) The evolving role of mass spectrometry in cancer biomarker discovery. *Cancer Biol Ther* **8:** 1083-1094

Wang Q, Chaerkady R, Wu J, Hwang HJ, Papadopoulos N, Kopelovich L, Maitra A, Matthaei H, Eshleman JR, Hruban RH, Kinzler KW, Pandey A, Vogelstein B (2011a) Mutant proteins as cancer-specific biomarkers. *Proc Natl Acad Sci U S A* **108:** 2444-2449

Wang Q, Rozelle AL, Lepus CM, Scanzello CR, Song JJ, Larsen DM, Crish JF, Bebek G, Ritter SY, Lindstrom TM, Hwang I, Wong HH, Punzi L, Encarnacion A, Shamloo M, Goodman SB, Wyss-Coray T, Goldring SR, Banda NK, Thurman JM et al (2011b) Identification of a central role for complement in osteoarthritis. *Nat Med* **17:** 1674-1679

Washburn MP, Wolters D, Yates JR (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **19:** 242-247

Weiss RB, Atkins JF (2011) Molecular biology. Translation goes global. *Science* **334:** 1509-1510

Wenger CD, Lee MV, Hebert AS, McAlister GC, Phanstiel DH, Westphall MS, Coon JJ (2011) Gas-phase purification enables accurate, multiplexed proteome quantification with isobaric tagging. *Nat Methods* **8:** 933-935

Whiteaker JR, Lin C, Kennedy J, Hou L, Trute M, Sokal I, Yan P, Schoenherr RM, Zhao L, Voytovich UJ, Kelly-Spratt KS, Krasnoselsky A, Gafken PR, Hogan JM, Jones LA, Wang P, Amon L, Chodosh LA, Nelson PS, McIntosh MW et al (2011) A targeted proteomics-based pipeline for verification of biomarkers in plasma. *Nat Biotechnol* **29:** 625-634

Wiese S, Reidegeld KA, Meyer HE, Warscheid B (2007) Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research. *Proteomics* **7:** 340-350

Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, Mathieson T, Lemeer S, Schnatbaum K, Reimer U, Wenschuh H, Mollenhauer M, Slotta-Huspenina J, Boese JH, Bantscheff M, Gerstmair A et al (2014) Mass-spectrometry-based draft of the human proteome. *Nature* **509:** 582-587

Williams KE, Miroshnychenko O, Johansen EB, Niles RK, Sundaram R, Kannan K, Albertolle M, Zhou Y, Prasad N, Drake PM, Giudice LC, Hall SC, Witkowska HE, Buck Louis GM, Fisher SJ (2015) Urine, peritoneal fluid and omental fat proteomes of reproductive age women: Endometriosis-related changes and associations with endocrine disrupting chemicals. *J Proteomics* **113:** 194-205

Wilm M (2011) Principles of electrospray ionization. *Mol Cell Proteomics* **10:** M111.009407

Wiśniewski JR, Nagaraj N, Zougman A, Gnad F, Mann M (2010) Brain phosphoproteome obtained by a FASP-based method reveals plasma membrane protein topology. *J Proteome Res* **9:** 3280-3289

Woods AG, Sokolowska I, Ngounou Wetie AG, Wormwood K, Aslebagh R, Patel S, Darie CC (2014) Mass spectrometry for proteomics-based investigation. *Adv Exp Med Biol* **806:** 1-32

Wölfler MM, Meinhold-Heerlein IM, Söhngen L, Rath W, Knüchel R, Neulen J, Maass N, Henkel C (2011) Two-dimensional gel electrophoresis in peritoneal fluid samples identifies differential protein regulation in patients suffering from peritoneal or ovarian endometriosis. *Fertil Steril* **95:** 2764-2768

Yamana R, Iwasaki M, Wakabayashi M, Nakagawa M, Yamanaka S, Ishihama Y (2013) Rapid and deep profiling of human induced pluripotent stem cell proteome by one-shot NanoLC-MS/MS analysis with meter-scale monolithic silica columns. *J Proteome Res* **12:** 214-221

Yan W, Lee H, Deutsch EW, Lazaro CA, Tang W, Chen E, Fausto N, Katze MG, Aebersold R (2004) A dataset of human liver proteins identified by protein profiling via isotope-coded affinity tag (ICAT) and tandem mass spectrometry. *Mol Cell Proteomics* **3:** 1039-1041

Yates JR, Cociorva D, Liao L, Zabrouskov V (2006) Performance of a linear ion trap-Orbitrap hybrid for peptide analysis. *Anal Chem* **78:** 493-500

Yau YY, Duo X, Leong RW, Wasinger VC (2015) Reverse-polynomial dilution calibration methodology extends lower limit of quantification and reduces relative residual error in targeted peptide measurements in blood plasma. *Mol Cell Proteomics* **14:** 441-454

Yin HR, Zhang L, Xie LQ, Huang LY, Xu Y, Cai SJ, Yang PY, Lu HJ (2013) Hyperplex-MRM: a hybrid multiple reaction monitoring method using mTRAQ/iTRAQ labeling for multiplex absolute quantification of human colorectal cancer biomarker. *J Proteome Res* **12:** 3912-3919

Young VJ, Brown JK, Maybin J, Saunders PT, Duncan WC, Horne AW (2014) Transforming growth factor-β induced Warburg-like metabolic reprogramming may underpin the development of peritoneal endometriosis. *J Clin Endocrinol Metab* **99:** 3450-3459

Yu D, Hung MC (2000) Overexpression of ErbB2 in cancer and ErbB2-targeting strategies. *Oncogene* **19:** 6115-6121

Zhang B, VerBerkmoes NC, Langston MA, Uberbacher E, Hettich RL, Samatova NF (2006a) Detecting differential and correlated protein expression in label-free shotgun proteomics. *J Proteome Res* **5:** 2909-2918

Zhang H, Liu AY, Loriaux P, Wollscheid B, Zhou Y, Watts JD, Aebersold R (2007) Mass spectrometric detection of tissue proteins in plasma. *Mol Cell Proteomics* **6:** 64-71

Zhang H, Liu Q, Zimmerman LJ, Ham AJ, Slebos RJ, Rahman J, Kikuchi T, Massion PP, Carbone DP, Billheimer D, Liebler DC (2011) Methods for peptide and protein quantitation by liquid chromatography-multiple reaction monitoring mass spectrometry. *Mol Cell Proteomics* **10:** M110.006593

Zhang H, Niu Y, Feng J, Guo H, Ye X, Cui H (2006b) Use of proteomic analysis of endometriosis to identify different protein expression in patients with endometriosis versus normal controls. *Fertil Steril* **86:** 274-282

Zhang R, Sioma CS, Wang S, Regnier FE (2001) Fractionation of isotopically labeled peptides in quantitative proteomics. *Anal Chem* **73:** 5142-5149

Zhang Y, Wen Z, Washburn MP, Florens L (2010) Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins. *Anal Chem* **82:** 2272-2281

Zhang Z, Bast RC, Yu Y, Li J, Sokoll LJ, Rai AJ, Rosenzweig JM, Cameron B, Wang YY, Meng XY, Berchuck A, Van Haaften-Day C, Hacker NF, de Bruijn HW, van der Zee AG, Jacobs IJ, Fung ET, Chan DW (2004) Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. *Cancer Res* **64:** 5882-5890

Zhao Y, Liu YN, Li Y, Tian L, Ye X, Cui H, Chang XH (2015) Identification of biomarkers for endometriosis using clinical proteomics. *Chin Med J (Engl)* **128:** 520-527

Zheng N, Pan C, Liu W (2011) New serum biomarkers for detection of endometriosis using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *J Int Med Res* **39:** 1184-1192

Zhou W, Liotta LA, Petricoin EF (2012) The spectra count label-free quantitation in cancer proteomics. *Cancer Genomics Proteomics* **9:** 135-142

Zhou Y, Shan Y, Zhang L, Zhang Y (2014) Recent advances in stable isotope labeling based techniques for proteome relative quantification. *J Chromatogr A* **1365:** 1-11

Zhou Y, Wong JM, Mabrouk OS, Kennedy RT (2015) Reducing adsorption to improve recovery and in vivo detection of neuropeptides by microdialysis with LC-MS. *Anal Chem* **87:** 9802-9809

Zhu L, Brown WC, Cai Q, Krust A, Chambon P, McGuinness OP, Stafford JM (2013) Estrogen treatment after ovariectomy protects against fatty liver and may improve pathway-selective insulin resistance. *Diabetes* **62:** 424-434

Zolotarjova N, Mrozinski P, Chen H, Martosella J (2008) Combination of affinity depletion of abundant proteins and reversed-phase fractionation in proteomic analysis of human plasma/serum. *J Chromatogr A* **1189:** 332-338

Zubarev RA (2013) The challenge of the proteome dynamic range and its implications for in-depth proteomics. *Proteomics* **13:** 723-726

Zubarev RA, Kelleher NL, McLafferty FW (1998) Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process. *Journal of the American Chemical Society* **120:** 3265–3266

Zuvich RL, Armstrong LL, Bielinski SJ, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, de Andrade M, Doheny KF, Haines JL, Hayes MG, Jarvik GP, Jiang L, Kullo IJ, Li R, Ling H, Manolio TA, Matsumoto ME, McCarty CA, McDavid AN et al (2011) Pitfalls of merging GWAS data: lessons learned in the eMERGE network and quality control procedures to maintain high data quality. *Genet Epidemiol* **35:** 887-898