PhD-FSTC-2015-45
The Faculty of Sciences, Technology and Communication

DISSERTATION

Defense held on 21/09/2015 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN CHIMIE

by

Stéphane TREVISIOL
Born on 16th July 1986 in Poissy, (France)

DEVELOPMENT OF ALTERNATIVE PROTEOMIC METHODS FOR LUNG CANCER BIOMARKER EVALUATION

Dissertation defense committee

Dr. Bruno Domon, dissertation supervisor
Luxembourg Institute of Health
Professor, Université du Luxembourg

Dr. Alain Van Dorsselaer
Professor, Institut Pluridisciplinaire Hubert Curien, Strasbourg

Dr. Eric Tschirhart, Chairman
Professor, Université du Luxembourg

Dr. Lennart Martens
Professor, Technische Universität Darmstadt

Dr. Iris Behrmann, Vice Chairman
Professor, Université du Luxembourg

UNIVERSITÉ DU LUXEMBOURG

Comité d’encadrement de thèse (CET)

Prof. Dr. Alain Van Dorsselaer

Prof. Dr. Eric Tschirhart

Prof. Dr. Bruno Domom

A mes grands-parents

TABLE OF CONTENTS

Acknowledgements

First of all, I thank the members of my thesis committee (CET), Prof. Van Dorsselaer and Prof. Tschirhart who have accepted to follow my PhD project. I am very grateful to them for their helpful advices. I want to thank my supervisor Bruno Domon who gave me the opportunity to carry out my PhD project at the Luxembourg Clinical Proteomics Center and for his guidance and support during these four years.

I'm very grateful to Sébastien, Daniel and Elodie, who taught me a lot during all my training with their patience. I thank our bioinformatician Sang Yoon for his computing contribution to my project. I also want to thank Dr. Guy Berchem, the PPM Program and the Integrated BioBank of Luxembourg for providing the clinical plasma samples. I'm very grateful to Jan for his reviews especially for my PhD thesis manuscript. I thank Yeoun Jin for her advice and support during my PhD project. I also want to thank Lizianne for her help in all the administrative part.

I thank all the members of the laboratory for their daily help, support and useful discussions. A special thanks to my PhD student colleagues Nina, Adèle, Lina and Daniela with whom I shared the PhD student life.

Last but not least, I am grateful for the financial support provided by the Fond National de la Recherche (FNR), for an AFR (n° 1194914) as well as the PEARL program (to BD, CPLI) which provided an access to a state-of-the-art platform.

List of abbreviations

aa: amino acid

AAA: quantitative amino acid analysis

ASMS: American Society for Mass Spectrometry

AUC: area under the curve

BLAST: basic local alignment search tool

CAD: collision activated dissociation

CE: collision energy

cICAT: carbon isotope-coded affinity tag

CID: collision induced dissociation

CIP: calibrated isotopically labeled peptide

CPP: concatenated polypeptides

CT-scan: computerized tomography scan

CV: coefficient of variation

DDA: data-dependent acquisition

DNA: deoxyribonucleic acid

DTT: dithiothreitol

EGF: epidermal growth factor

ESI: electrospray ionization

ETD: electron transfer dissociation

eV: electron volt

FDR: false discovery rate

FNR: Fond National de la Recherche

HCD: Higher-energy C-trap dissociation

HDL: high density lipoprotein

HF: high frequency

HPLC: high performance liquid chromatography

HUPO: Human Proteome Organization

IAA: iodoacetamide

IARC: International Agency for Research on Cancer

IBBL: Integrated Biobank of Luxembourg

ICAT: isotope-coded affinity tag

ISP: isotopically-labeled signature peptide

iTRAQ: isobaric tags for relative and accurate quantification

LC: liquid chromatography

LDL: low density lipoprotein

LOD: limit of detection

LOQ: limit of quantification

m/z: mass/charge ratio

MALDI: matrix-assisted laser desorption ionization

MS/MS: tandem mass spectrometry

MS: mass spectrometry

MudPIT: multidimensional liquid chromatography separation

nCE: normalized collision energy

NIH: National Institutes of Health

NSCLC: non-small cell lung cancer

PET: positron emission tomography

PPM: Partnership for Personalized Medicine

PRM: parallel reaction monitoring

PSAQ: protein standard absolute quantification

PTM: post translational modification

Q: quadrupole

QconCAT: isotope labeled concatamer of proteotypic peptides

RP: reporter peptide

RRF: relative response factor

Rt: retention time

SAA: serum amyloid A

SCLC: small cell lung carcinoma

SIL: stable isotope-labeled

SILAC: stable isotope labeling with amino acids in cell culture

SPE: solid phase extraction

SRM: selected reaction monitoring

SSRcalc: sequence-specific retention calculator

TMT: tandem mass tags

TNM: tumor nodule metastasis

TOF: time of flight

UPS1: universal protein standard

WHO: World Health Organization

XIC: extracted ion chromatogram

Summary

Lung cancer is a serious public health problem, killing millions of people around the world every year. It is the third most common cancer in Europe after colorectal cancer and breast cancer, but the deadliest since it is often detected too late. Sensitive and reliable methods for early stage detection and drug response prediction could significantly improve the therapeutic treatment and survival rate of lung cancer patients. Although tremendous efforts in biomedical research have been made over the last decades to develop novel biomarkers for cancer diagnosis and treatment response, no reliable and specific panels of biomarkers are available for early lung cancer diagnosis. During the past decade mass spectrometry (MS) based proteomics has become an important approach to biological and clinical investigations. Most MS-based protein quantification approaches generally involve trypsin digestion of the endogenous proteins in biological samples, followed by a targeted liquid chromatography-mass spectrometry (LC-MS) analysis of signature peptides indicative of the proteins of interest. Trypsin cleaves proteins C-terminal of lysine and arginine residues generating peptides that are well suited for LC-MS/MS analysis. However, despite the undeniable advantages of trypsin, the enzyme is not optimal for all types of proteomics studies due to an uneven distribution of the cleavage sites in the human proteome. During a tryptic digestion also very short non-specific peptides (< 6 amino acids) are produced resulting in a loss of sequence protein coverage. This can be problematic as the missing peptide parts could contain crucial information for disease understanding, such as specific amino acid mutations or post-translational modifications of amino acid residues. Moreover, the generation of large numbers of short peptides results in an increased background complexity which is a main limiting factor in quantification experiments as it reduces the selectivity of the measurements, which in turn affects the sensitivity of the experiments. Isotope dilution strategies are frequently employed to achieve accurate quantification of the proteins of interest using calibrated isotopically labeled peptides as internal standards. In quantitative assays, which involve the use of stable isotope-labeled standards, the reliability and the accuracy of the experiments is principally dependent on the quality of the standards. Different factors can affect the standard concentration between the synthesis and its utilization, such as incomplete solubilization, non-specific adsorption to hydrophobic surfaces or aggregation. To ensure an accurate quantification the internal standards need to be verified immediately before analysis. The purpose of this PhD thesis was to develop new proteomic approaches to remedy these limitations by using, instead of trypsin, alternative enzymes and by developing a quantification approach based on concatenated polypeptide standards containing a cleavable reporter peptide for accurate quantification.

Outline of the thesis

The aim of the PhD thesis was to develop quantitative proteomic strategies, in the context of lung cancer biomarker evaluation, to overcome the limitations of the classical proteomic approach. The thesis is divided in six sections, an introduction, three chapters, a material & methods section, and the conclusion.

The introduction provides the background focused on topics related to this research project namely lung cancer, biomarkers and MS based-proteomics.

The first chapter presents an accurate quantification strategy involving the use of a new type of concatenated polypeptide as internal standard. It contains a cleavable reporter peptide allowing the systematic calibration/recalibration of the standard just before performing quantitative analyses. The design of the standard, the description of the different workflows and their application for the quantification of protein biomarker candidates in lung cancer plasma samples are presented.

In the second chapter, an assessment of alternative proteolytic enzymes to trypsin is conducted in order to enhance proteomics analyses. First, the capabilities of alternative enzymes to reduce the loss of proteome sequence coverage during tryptic digestion and to decrease the peptide density was evaluated. Second, the surrogate peptides were empirically selected to target eight non-small cell lung cancer (NSCLC) biomarker candidates and tested upon proteolysis of the corresponding recombinant proteins with five different proteolytic enzymes. Finally, the fragmentation behaviors of the corresponding peptides were investigated to optimize the collision energy in order to improve the sensitivity of targeted quantitative experiments.

The third chapter describes the application of the cleavable reporter peptide strategy developed in the first chapter to the quantification of four NSCLC protein biomarker candidates, measured in twenty four clinical plasma samples. The samples were digested in parallel with Lys-C and trypsin targeting common peptides between both digestions in order to establish that quantitative analyses performed with Lys-C can result in, at least, a similar analytical performance as obtained with a standard tryptic digestion.

The two last parts of the manuscript describe the experimental part (material & methods) and the conclusion & outlook of the performed research.

Introduction

In 2014, the International Agency for Research on Cancer (IARC), a specialized department of the World Health Organization (WHO), mandated to conduct and coordinate research on cancer, drew a dark picture on cancer development in the world. In its latest report, *World Cancer Report 2014* [1]*,* it highlights the increase of the worldwide burden due to cancer. In 2012, cancer was the leading cause of death (8.2 million) exceeding ischemic heart diseases (7.4 million) and stroke (6.5 million). Among the number of cancer types, lung cancer was the most frequent and deadliest with 1.8 million cases and with 1.6 million deaths. It was followed by breast (incidence: 1.7 million, deaths 0.5 million) and colorectal (incidence: 1.4 million, deaths 0.7 million) cancers. In this report, predictions of the evolution of cancer incidence for the coming 20 years were also presented. The annual number of new cancer cases is expected to exceed 22 million people and the yearly number of deaths to increase until 13 million.

1. Lung cancer

Cancer is in fact a generic term designating diseases characterized by an abnormal proliferation and an accumulation of tumor cells in healthy tissues inducing physiological disturbances, ultimately resulting in the death of the living organism [2]. Many types of cancer start from one cell which has acquired several characteristics to become a tumor cell by modifications and alterations of its genetic program [3]. Tumorigenesis is a multi-step process in which tumor cells acquire different hallmarks such as evading apoptosis which is a process initiating cell-death for example to eliminate damaged cells [4]. Lung cancer is the most prevalent type of cancer in the world. The high proportion of lung cancer is directly attributable to tobacco consumption. Indeed, 90% of the lung cancer cases are directly related to smoking and passive smoking [5]. In tobacco smoke, more than 5000 chemical components were identified, and at least 60 of them were recognized as carcinogens [6]. Other environmental causes also contribute to the risk of developing lung cancer such as exposure to carcinogens (asbestos, radon or chemicals) [7] or air pollution (fine particulates) [8]. There are four main different histological types of lung cancer [9]. Squamous-cell carcinoma (50%) is the most common form of lung cancer and is mainly attributed to the agents contained in cigarette smoke. This slow growing tumor originates as a squamous metaplasia at a main bronchus and can be removed by surgery. Adenocarcinoma (20%) develops at the lung periphery and only very rarely shows observable signs. Since it is found among smokers and nonsmokers, this tumor type is not directly a consequence of cigarette smoking. Moreover, adenocarcinomas are the most commonly diagnosed type of lung cancer among non-smokers [10]. Large cell carcinoma (10%) is the development of large undifferentiated cells with polymorphic nuclei (they have adenocarcinoma and squamous components) at the Hilar region of the lung or at the lung periphery. This form of cancer has a poor prognosis as it usually has already began to spread throughout the rest of the body at the time of diagnosis. Small cell carcinoma (20%) is the most malignant form of lung cancer. These cells originate from the bronchial epithelium in the Hilar region. This type of cancer is strongly associated with smoking, and has the worst prognosis because it is often only diagnosed at an advanced metastatic stage. According to the morphology of the lung cancer tumor cells and their response to treatment, lung cancer cases are grouped into two broad categories which are non-small cell lung carcinoma NSCLC (squamous cell carcinoma, adenocarcinoma, large cell carcinoma) and small cell lung carcinoma SCLC.

Generally lung cancer is diagnosed when the first symptoms of the disease are a worsening of a persistent cough or chest discomfort. The first steps in the diagnostic process of lung cancer consist in a clinical examination followed by chest radiography. However, only a biopsy allows a direct and confirmed diagnosis including the information on the type of cancer. When the disease is diagnosed, several types of exams exist to assess the progression degree. These include mediastinoscopy (to determine whether cancer cells have invaded the lymph nodes of the mediastinum), bone scintigraphy (to search for cancer cells localized in bones), CT-scans to estimate the size of the tumor, and the degree of invasiveness, and PET scans to visualize the cancer cells [11]. The staging of cancer is codified by the Tumor Nodule Metastasis (TNM) classification of malignant tumors [12]. Using this nomenclature, the stage of the cancer can be determined, from stage 0 to stage IV, based on the degree of dispersion of the disease allowing oncologists to choose the most appropriate treatment.

Unlike breast and colorectal cancers, no early detection tests are available for lung cancer. To successfully establish the diagnosis of lung cancer, the patient undergoes many tests that are invasive, cumbersome and expensive. This reflects the screening limitations for lung cancer today and highlights the need for an easy and non-invasive test for early detection. The use of biomarkers, reflecting changes occurring during the tumorigenesis process, is a promising approach for lung cancer early detection.

2. Biomarkers

The American National Institutes of Health (NIH) defines a biomarker as “a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention” [13]. In the case of cancer, tumor markers are going to be biological indicators of molecular changes that occur during malignancy. They generally reflect genetic alterations such as mutations in genes, in their products or in their post-translational modifications (PTMs) [14]. Molecular and functional characterization of all these changes and their effects can result in the establishment of a molecular description of tumors at a specific stage of tumor development.

The use of biomarkers can help physicians at the different stages of cancer curation *i.e.*, identification, characterization, treatment and monitoring (Figure 1).

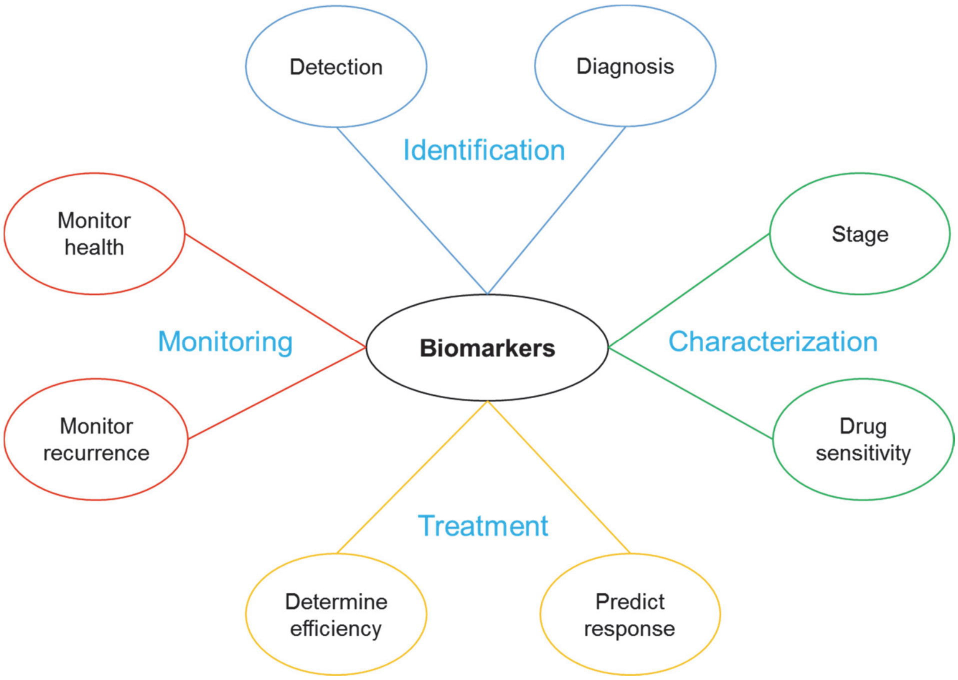


Figure 1: Diagram of the different uses of biomarkers (figure from US Biomarkers Inc.).

An “ideal” biomarker can be defined by several characteristics. Its measurement should be easy and non-invasive, and the cost of the method should be relatively low, it should also be consistent across all the population and it should present high specificity and sensitivity [15]. As described in Figure 2, the sensitivity is the ability of the biomarker to properly identify all people who are not affected by this particular cancer (false-positives) and the specificity is the ability of a biomarker to properly identify all patients affected with a particular cancer (false-negatives). For clinical application, a biomarker needs to reach at least 90% of sensitivity and specificity [16].

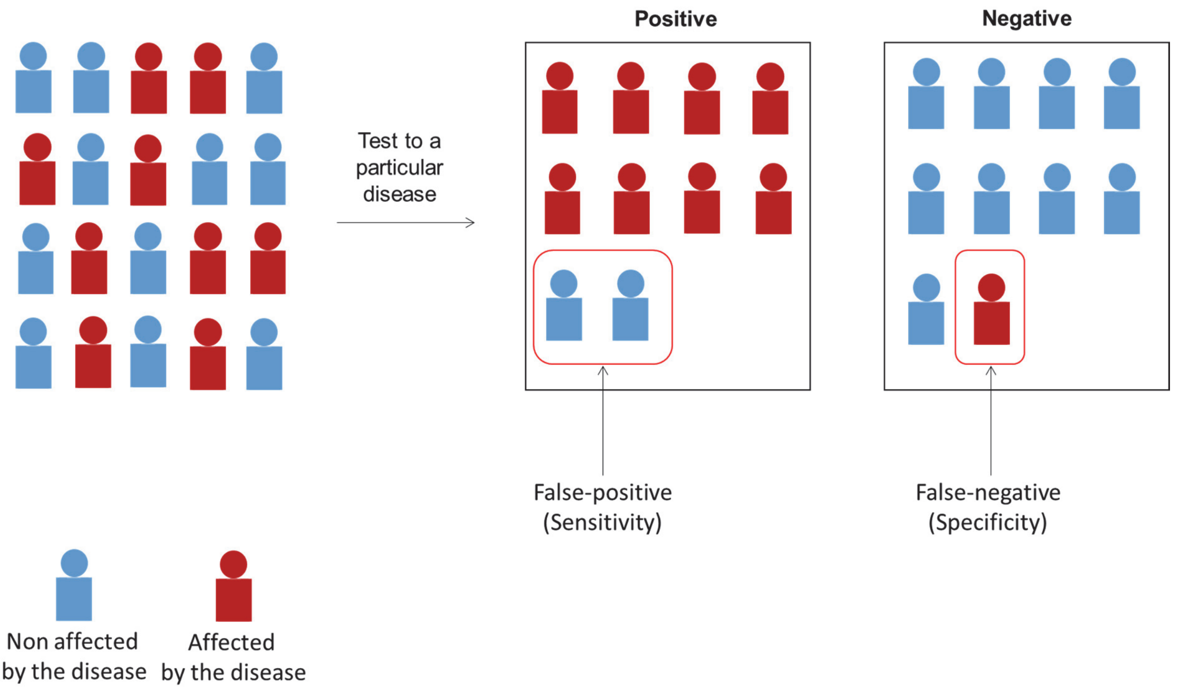

Positive
Negative
Test to a particular disease
False-positive (Sensitivity)
False-negative (Specificity)
Non affected by the disease
Affected by the disease

Figure 2: Illustration of the definitions of sensitivity and specificity.

Currently some biomarkers have been characterized and are already used in a clinical context. Some of them are associated with a single type of cancer; *e.g.* the carbohydrate antigen 15-3 used to assess treatment efficiency of breast cancer [17]. Other markers are associated with several types of cancer such as the carcinoembryonic antigen used to evaluate the spread of colorectal cancer and also to assess the treatment response for breast cancer [18]. Today, most identified cancer biomarker candidates do not have enough sensitivity and specificity to be used for cancer screening but it has been shown that specificity and sensitivity of a single biomarker can be improved by combining individual biomarkers into a biomarker panel [19].

The pipeline for the identification of new biomarkers is a long process which can be split in three distinct phases: discovery, evaluation and pre-clinical (Figure 3).

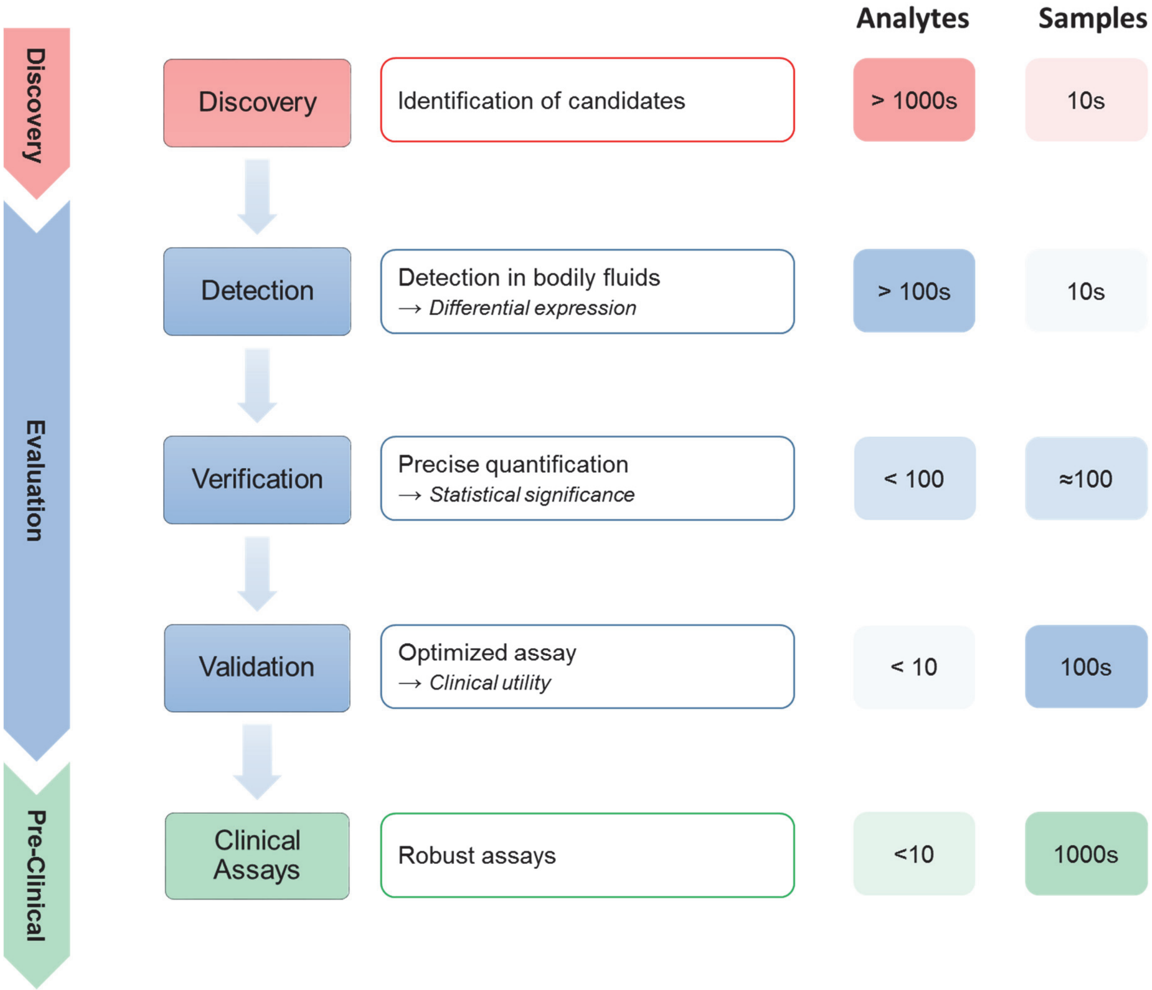
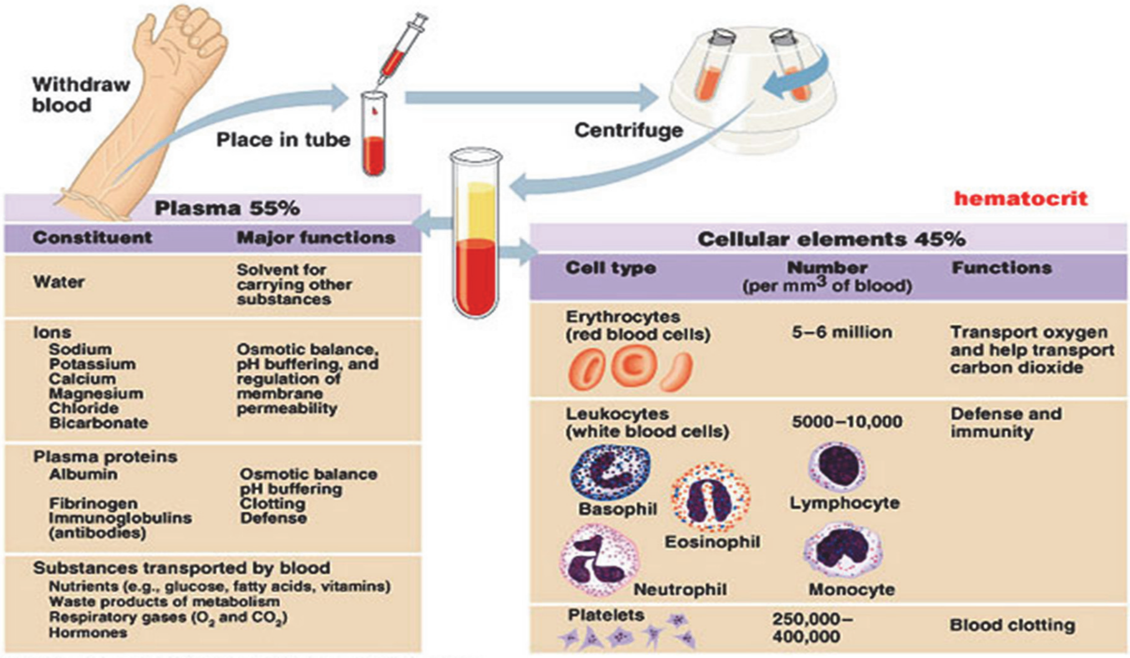


Figure 3: Workflow for the identification of new biomarker candidates (adapted from Domon et *al.,* Proteomics Clinical Applications., 2015; 9: 423–431).

The purpose of the initial discovery phase is to identify protein biomarker candidates which are differentially expressed between diseased and healthy individuals. An unbiased semi-quantitative analysis can be performed on different types of samples such as proximal fluids, cell line supernatants, animal model plasma or human plasma [20]. In the discovery phase thousands of proteins are selected for screening as candidate biomarkers in a limited number of biological samples. The second phase of the biomarker workflow is the evaluation [21]. The objective of this step is first to detect in a limited number of bodily fluid samples (10s) the hundreds of protein candidates which were selected during the discovery phase to confirm their differential abundance. In a second step, the precise quantification of a subset of differentially expressed biomarker candidates (<100) is performed in a larger number of samples (≈100) to verify their sensitivity and specificity [22]. Finally, the most promising markers (<10) which present a high degree of sensitivity and specificity are measured in a set of hundreds of biological samples using optimized analytical conditions to document the quantification performance of the assay (the linearity range, the limits of detection, and quantification, precision and accuracy of the measurements). The limited number of potential biomarkers (<10) which succeed to pass the evaluation phase can be used as targets for preliminary clinical assays commonly performed using gold standard immunoassay tests on thousands of human samples reflecting the extent of variability in the human population [23].

The ultimate goal of all the studies for the identification and the validation of a new biomarker is at term to develop a simple non-invasive test able to improve the care of patients. Blood is a vital fluid easily accessible in constant interaction with all the parts of the body. Its dynamic composition reflects the physiological or pathological states of a person making it an obvious choice for new biomarker investigations [24]. Blood is composed of two main parts which are easily separable by centrifugation: the sediment (red color) and the plasma (yellow color) representing 45% and 55% of the total blood volume, respectively. The sediment is constituted of all the blood cells (erythrocytes, leukocytes and thrombocytes, *etc.*). Plasma is the aqueous solution in which blood cells are in suspension. It is mainly composed of proteins, hormones, sugars and lipids (Figure 4).



Plasma 55%

Constituent	Major functions
Water	Solvent for carrying other substances
Ions Sodium Potassium Calcium Magnesium Chloride Bicarbonate	Osmotic balance, pH buffering, and regulation of membrane permeability
Plasma proteins Albumin Fibrinogen Immunoglobulins (antibodies)	Osmotic balance pH buffering Clotting Defense
Substances transported by blood Nutrients (e.g., glucose, fatty acids, vitamins) Waste products of metabolism Respiratory gases (O_2 and CO_2) Hormones	

Cellular elements 45%

Cell type	Number (per mm^3 of blood)	Functions
Erythrocytes (red blood cells)	5–6 million	Transport oxygen and help transport carbon dioxide
Leukocytes (white blood cells)	5000–10,000	Defense and immunity
Platelets	250,000–400,000	Blood clotting

Figure 4: Illustration of blood composition (figure from Cummings, Pearson Education Inc.)

Plasma analyses have been used in medicine since decades for the diagnosis of many diseases such as for cardiovascular diseases with the determination of high density lipoprotein (HDL) and low density lipoprotein (LDL) levels or for diabetes with the determination of glucose concentration. Human plasma has been studied for a long time and is daily used for clinical analyses but remains unfully characterized mainly due to its high degree of complexity [20]. Indeed, the human plasma proteome contains more than ten thousands proteins (10546 were registered in the *Plasma Proteome Database* 2014 [25]) which are present in a large range of concentrations. The dynamic range encompasses up to 12 orders of magnitude and with the presence of very abundant proteins such as serum albumin (35-50 mg/mL) and immunoglobulins (5-18 mg/mL) it becomes very challenging to analyze very low abundant proteins that could be potential biomarkers. The plasma proteome can be categorized in three overlapping groups of proteins based on their abundances: classical plasma proteins (from 50 mg/mL to 10 ng/mL), tissue leakage products (from 500 ng/mL to 100 pg/mL) and interleukins/cytokines (from 1 ng/mL to 0.05 pg/mL) [26, 27] (Figure 5).

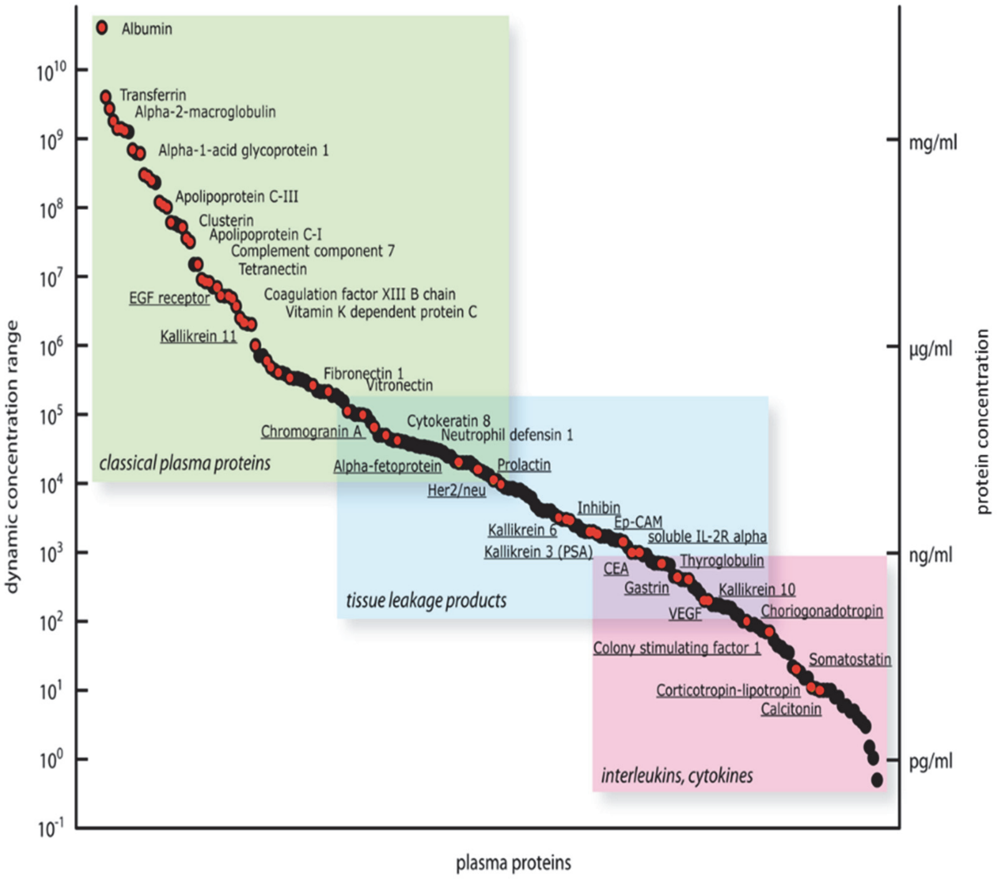


Figure 5: Representation of plasma protein concentrations distributed over 12 orders of magnitude (from Schiess et *al.,* Molecular Oncology, 2009; 3, 33–44).

Moreover, in the plasma proteome some proteins are also present in different isoforms and in addition substantial variations are also introduced by PTMs such as phosphorylation and glycosylation. Including all the variants and PTMs the total number of proteins in the plasma proteome should approach the million [28, 29].

The analysis of complex biological samples as human plasma remains very challenging. Identification and quantification of low abundant proteins such as biomarkers in clinical plasma samples require high-throughput analytical methods able to separate the different constituents over a wide dynamic range of concentrations, with high sensitivity and specificity [30]. Mass spectrometry based proteomics strategies involving an effective sample preparation, a liquid chromatography separation and a mass spectrometry detection have many assets to meet the challenge of the analysis of low abundant proteins in complex biological samples.

3. Mass spectrometry based proteomics

Proteomics is the large-scale comprehensive study aiming at the characterization of a specific proteome including identification and quantification of all the proteins, and determination of their expression levels, localization, interactions or post-translational modifications [31]. The proteomics terminology was coined in 1997 by merging two words “proteins” and “genomics” [32].

One of the most powerful tools for proteomic studies is mass spectrometry. This analytical method is based on the creation and the detection of charged compounds in a gas phase. A mass spectrometer is composed of three main elements, a source, a mass analyzer and a detector. In the source, molecules are ionized and transferred to the gas phase. Ions are then transmitted to the mass analyzer due to electromagnetic fields where they will be separated based on their mass-to-charge ratios (m/z). To finish, ions reach the detector generating an ion currant which will be subsequently converted into a mass spectrum [33]. A second analyzer can be present to perform tandem mass spectrometry in order to fragment ions giving access to structural information additionally to their mass. Different soft ionization techniques such as electrospray ionization (ESI) [34] and matrix-assisted laser desorption ionization (MALDI) [35] are commonly used to generate and to transfer into the gas phase intact biomolecules as peptides and proteins. Currently, mass spectrometry is in an outstanding position among all analytical methods for the identification and quantification of protein in complex biological samples such as biomarkers in bodily fluids, due to its high sensitivity, low limit of detection, acquisition speed and flexibility [36].

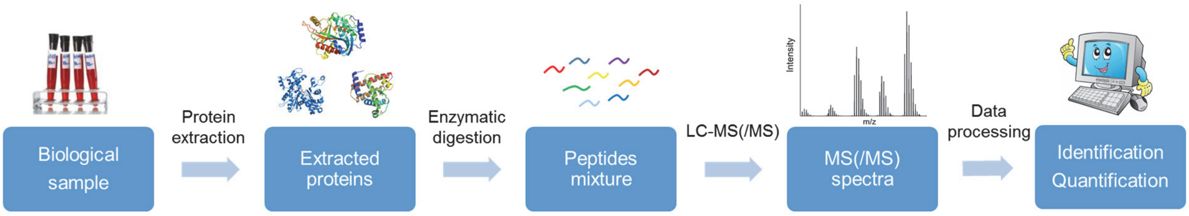


Figure 6: Workflow of bottom-up MS-based proteomics.

For the analysis of complex samples, the bottom-up proteomic strategy is mostly used. As presented in Figure 6, protein characterization is performed in four steps. Proteins of interest are first extracted from the biological samples. Afterwards, extracted proteins are enzymatically digested producing peptides which will be separated using liquid chromatography and analyzed by mass spectrometry generating MS and MS/MS spectra. Last, MS spectra are processed using informatics tools to perform peptide identification and quantification [37].

Protein extraction

Plasma is a promising source of protein biomarkers but due to the large dynamic range of concentrations, the analysis of low abundant proteins remains very challenging. In plasma, the ten most abundant proteins represent approximately 90% of the total protein amount [38]. One strategy to decrease plasma complexity is to perform a prefractionation by immunodepletion in order to selectively remove the most abundant proteins to reduce the dynamic range of concentrations [39]. Currently it is possible to deplete plasma up to 20 proteins [40]. Immunodepletion is based on the use of polyclonal antibodies fixed in an analytical column which will specifically retain a specific protein, their “antigen”. The plasma immunodepletion process for albumin and immunoglobulin proteins is presented in Figure 7.

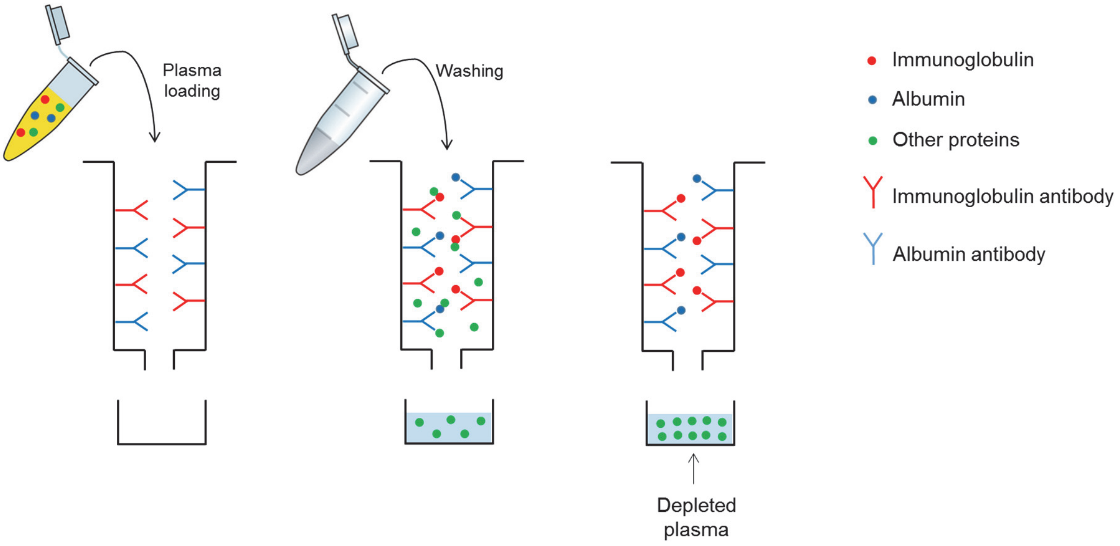


Figure 7: Plasma immunodepletion of albumin and immunoglobulin proteins.

First the plasma sample is loaded into the immunodepletion cartridge where albumin and immunoglobulin will be binding to their specific antibody whereas other proteins will pass through. After this, the cartridge is washed to detach proteins which are unspecifically bounded to the albumin and immunoglobulin antibodies or blocked into the cartridge. The reunion of the two flow-throughs corresponds to the depleted plasma (immunoglobulin/albumin) which can be further enzymatically digested.

Enzymatic digestion

In the bottom-up strategy, protein characterization is based on an indirect measurement of proteins by the analysis of surrogate peptides generated during an enzymatic digestion process. Usually the proteolysis is performed with trypsin a well-known pancreatic enzyme fully active at pH~8 identified by Wilhelm Kühne in 1876, that efficiently cleaves proteins at the carboxylic side of lysine and arginine residues. If a proline is located at the C-terminal part of these residues the cleavage frequency is very low [41]. The specificity of trypsin is ensured via electrostatic interactions between the negative charge carried by the aspartic acid (Asp198) located at the bottom of the catalytic pocket with the positive charge of lysine and arginine residues present at pH 8. Trypsin is a serine protease characterized by a catalytic triad composed of three amino acids, aspartic acid (Asp102), histidine (His57) and serine (Ser195) which act the roles of acid, base and nucleophile, respectively, to hydrolyze the peptide backbone [42]. The carboxylate group of the aspartic acid is in interaction with the histidine via a hydrogen bond which has the effect to increase the pKa of its imidazole nitrogen from 7 to 12. At this point, the histidine is considered as a strong base able to deprotonate the alcohol group of the serine residue, therefore activating the enzyme. In the catalytic mechanism of serine proteases (Figure 8) [43], the first step consists in a nucleophile attack of the activated serine against the peptide backbone, which induces the of generation a first tetrahedral intermediate (reaction 1). Then, this intermediate is decomposed in an acyl-enzyme intermediate by rupture of the peptide backbone (reaction 2). Next, the cleaved peptide with a new N-terminal part is released by the acyl-enzyme intermediate and replaced with a molecule of water (reaction 3). Finally, the two last steps of the catalytic mechanism correspond to a saponification reaction between the ester group linked to the serine and a hydroxide anion generated after the deprotonation of the molecule of water by the histidine. Reaction 4 is the addition of a hydroxide anion to the ester. Reaction 5 is the elimination of the alkoxide group which regenerates the enzyme and releases the cleaved peptide with a new C-terminal part. At this point the enzyme is ready to catalyze a new cleavage cycle.

Trypsin is the most commonly used enzyme in bottom-up proteomics with which a high sequence protein coverage is obtained [44] but due to an uneven distribution of the cleavage sites in the human proteome very short non-specific peptides (< 6 amino acids) are also produced resulting in a loss of sequence protein coverage [45]. These missing peptides can be an issue especially in the context of biomarkers because they may contain important information such as PTMs or amino acid mutations [46]. Alternative enzymes to trypsin such as Lys-C, Glu-C, Asp-N and Arg-C which have different cleavage specificities can also be used to perform the proteolysis [47]. Lys-C is a serine protease like trypsin, active at pH 9, which specifically cleaves proteins after lysine residues. The catalytic triad of Lys-C is formed by the residues aspartic acid (Asp113), histidine (His57) and (Ser194). The attraction of lysine residues into the catalytic pocket is ensured via the negative charge carried by the aspartic acid (Asp225) [48]. Glu-C is also a serine protease active at pH 7.5-8 cleaving after acidic amino acids with a preference for glutamic acid. It cleaves 3000-fold faster after glutamic acid residues than after aspartic acid, independently of the buffer used [49]. The catalytic triad of Glu-C is composed of the residues aspartic acid (Asp93), histidine (His51) and serine (Ser169). Asp-N is a zinc metalloendopeptidase, active in the pH range 7.0 - 8.5, cleaving proteins before aspartic acid and cysteic acid residues. It also cleaves before glutamic acid residues at a slower rate. In the catalytic site an atom of zinc is bound with three histidine residues within the motif $H^{167}EXXH^{171}XXGXXH^{177}$. The catalytic mechanism of Asp-N (Annex 1) involves the glutamic acid (Glu168) [50]. Arg-C is a cysteine-activated protease active in the pH range 7.6 - 9 that cleaves proteins after arginine residues [51].

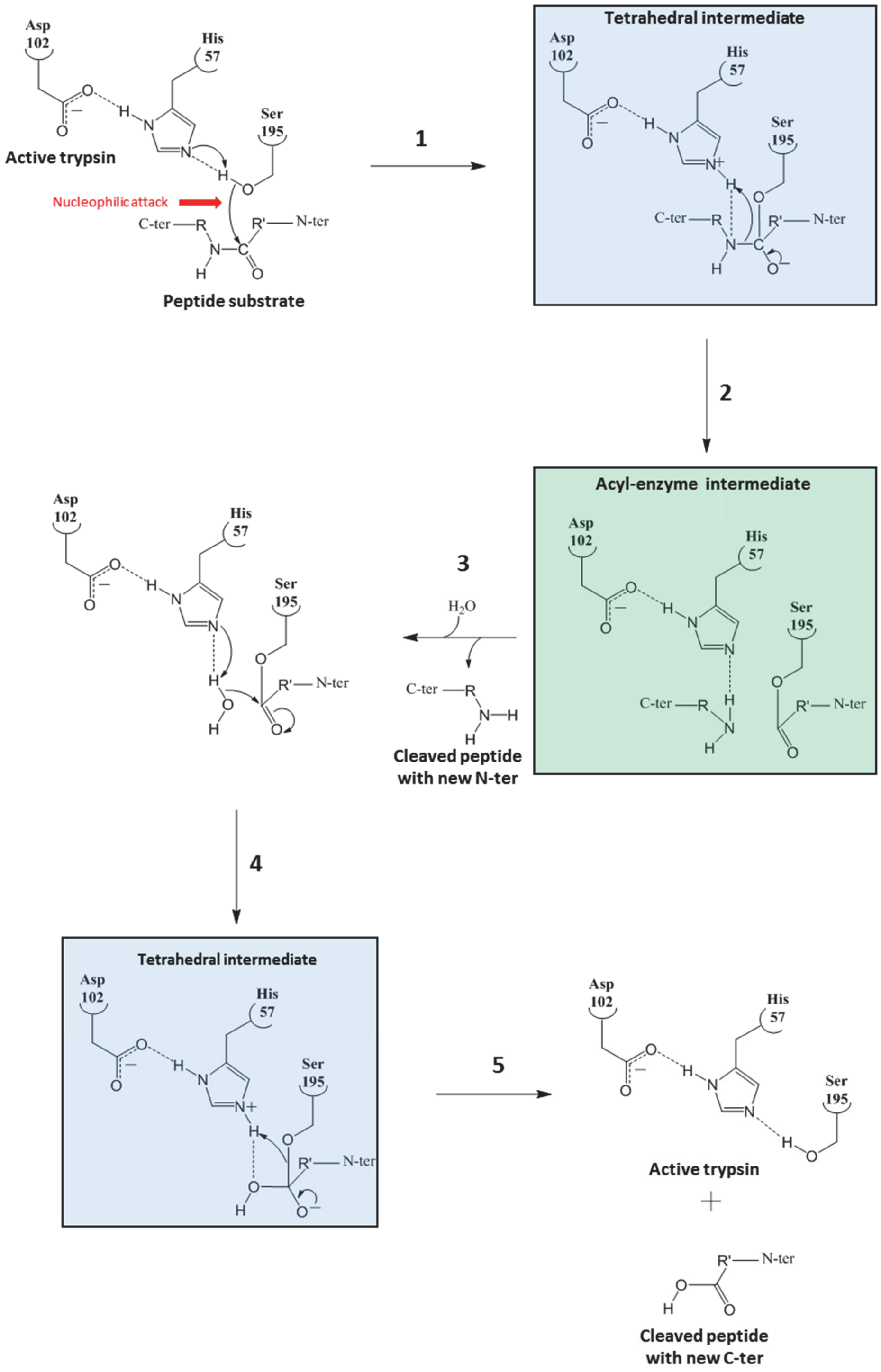


Figure 8: Serine proteases catalytic mechanism applied to trypsin (from Voet, Principles of Biochemistry, chapter 1).

The active site of the enzyme is composed of one histidine (His176) and one cysteine (Cys231). The attraction of arginine residues into the catalytic pocket is performed through the negative charge carried by the aspartic acid (Asp229). The activity of Arg-C is highly dependent on the cysteine thiol group of the enzyme, thus the presence of a reducing agent is required. The mechanism is presented in Annex 2 [52].

In most proteomics studies the use of trypsin has been preferred over other enzymes because it cleaves proteins after lysine and arginine residues, generating peptides with molecular masses ideally suited for the m/z range of quadrupole analyzers and put in C-terminal basic amino acids allowing an efficient fragmentation [53].

Peptide fragmentations

During the last decades mass spectrometry has proven to be an essential tool for protein sequencing especially after the development of tandem mass spectrometry by McLafferty in 1983 [54]. Tandem mass spectrometry experiments are characterized by at least two steps of mass analyses conjugated with a fragmentation stage in between. Tandem mass spectrometry methods can be designed in two different ways, in space or in time, based on the characteristics of the used instrumentation (Figure 9).

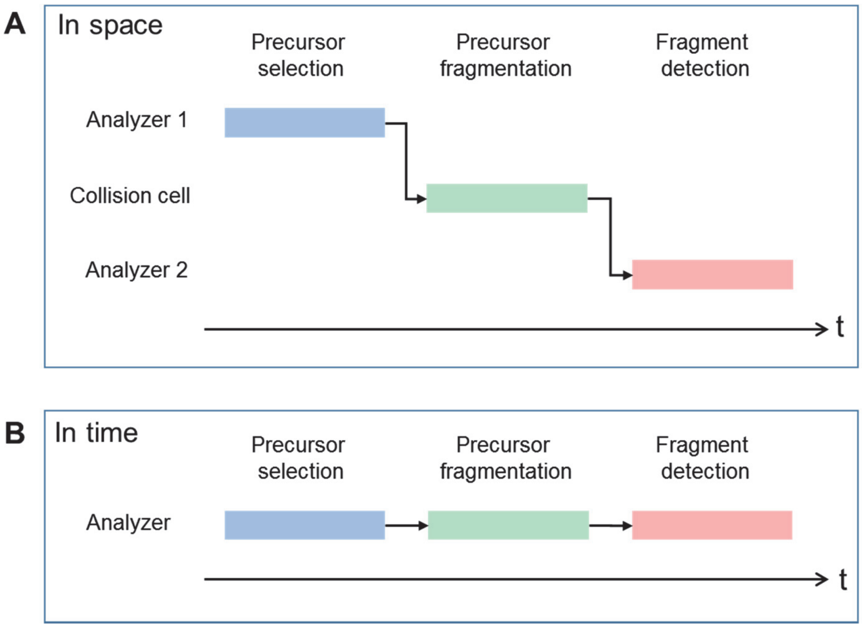


Figure 9: Illustration of the two different tandem mass spectrometry acquisition modes. In space (A) and in time (B) (adapted from Hoffmann, Mass Spectrometry Principles and Applications Third Ed.).

For tandem mass spectrometry in space (Figure 9A), generally the instruments possess at least two mass analyzers such as triple quadrupole [55], time of flight (TOF)-TOF [56], quadrupole-orbitrap [57, 58] or quadrupole-TOF [59] mass spectrometers. In the first analyzer, specific precursor ions are selected and isolated before they are transferred into a collision cell where they are fragmented; finally, the fragment ions are transferred to a second mass analyzer where their m/z values are measured. In tandem mass spectrometry performed in time (Figure 9B), all the steps of the experiment, as explained before for the tandem mass spectrometry in space, are performed at the same location but are separated in time. The selection of the precursor ion, its fragmentation and its detection are performed sequentially. Mass analyzers such as ion traps (*e.g.,* linear trap) [60] perform tandem mass spectrometry experiments in time, using the same analyzer for selection and fragmentation.

Nomenclature of the different peptide fragments

Fragmentation patterns of protonated peptides are a function of different parameters such as amino acid composition, size of the peptides, charge state of the precursors, amount of internal energy transferred, etc. In order to classify peptide fragments a nomenclature based on the location of the cleavage sites within the peptide bonds was proposed by Roepstorff and Fohlman [61] in 1984 and modified later by Biemann [62] in 1988 (Figure 10). On the main chain of peptide bonds the cleavage can occur on three types of chemical bonds $C_\alpha – C$, $C – N$ and $N – C_\alpha$ generating six types of fragments respectively a_i, b_i, c_i if the fragment includes the N-terminal part and x_{n-i}, y_{n-i}, z_{n-i} if the fragment includes the C-terminal part (n indicates the number of amino acids of the fragment and i the position starting from the N-terminal site).

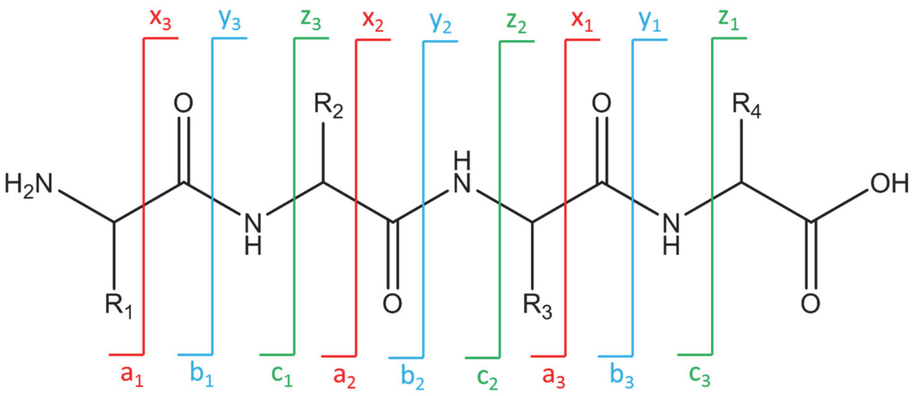


Figure 10: Peptide fragment ion nomenclature for a four amino acid peptide.

Consecutive fragmentations can occur on peptide fragments producing shorter fragments or internal fragments (fragments which have lost the C-terminal and N-terminal sites) if the second fragmentation is located on the main chain. If the second fragmentation occurred on the lateral chain of amino acids, the satellite ions d, v, w are produced from ions a, y, z, respectively [63] .

Collision induced dissociation (CID)

Collision induced dissociation (CID), also named collision activated dissociation (CAD), is one of the most commonly used fragmentation modes in proteomics [64]. In this activation mode, precursor ions are excited by collision with an inert gas (i.e. helium or argon) transferring a fraction of their translational energy into internal energy putting them into an excited state. CID is an ergodic activation method meaning that the process of dissociation of activated ions is slower than the process of random redistribution of energy. Thus, after activation the excess of internal energy is equally distributed by all the ion vibration modes so the first chemical bonds to cleave correspond to the weakest. In CID, two energetic regimes can be defined: high-collision energy (keV, electronic excitation) and low-collision energy (1-100 eV, vibrational excitation) based on the capabilities of the instrument used. Between high and low collision energy different fragmentation patterns are observed because fragmentation is a function of the available amount of energy.

Most proteomics experiments for peptide identification and quantification are performed in low-energy mode because mainly b ions, y ions and neutral loss (ammonia and water) are produced making the interpretation of MS/MS spectra more straightforward [65]. To describe the dissociation process of protonated peptides under low-collision energy activation the mobile proton model [66] was proposed by Wysocki in 1996. During the ionization process, peptide protonation can occur on various sites like terminal amino groups, amide groups (oxygen and nitrogen) and basic side chain groups. Based on the gas phase proton affinity of these groups, the protonation will take place first on the basic side chain groups (lysine, arginine and histidine) and second on the terminal amino groups which are thermodynamically more stable than the oxygen and nitrogen atoms of the amide groups. Upon excitation the internal energy of protonated peptides increases allowing the migration of not sequestered protons to a less favorable protonation site, the nitrogen amide group, which initiates peptide dissociation generating y- and b- ions. Figure 11 shows the charge-directed fragmentation mechanism of a tryptic peptide using the model of the mobile proton.

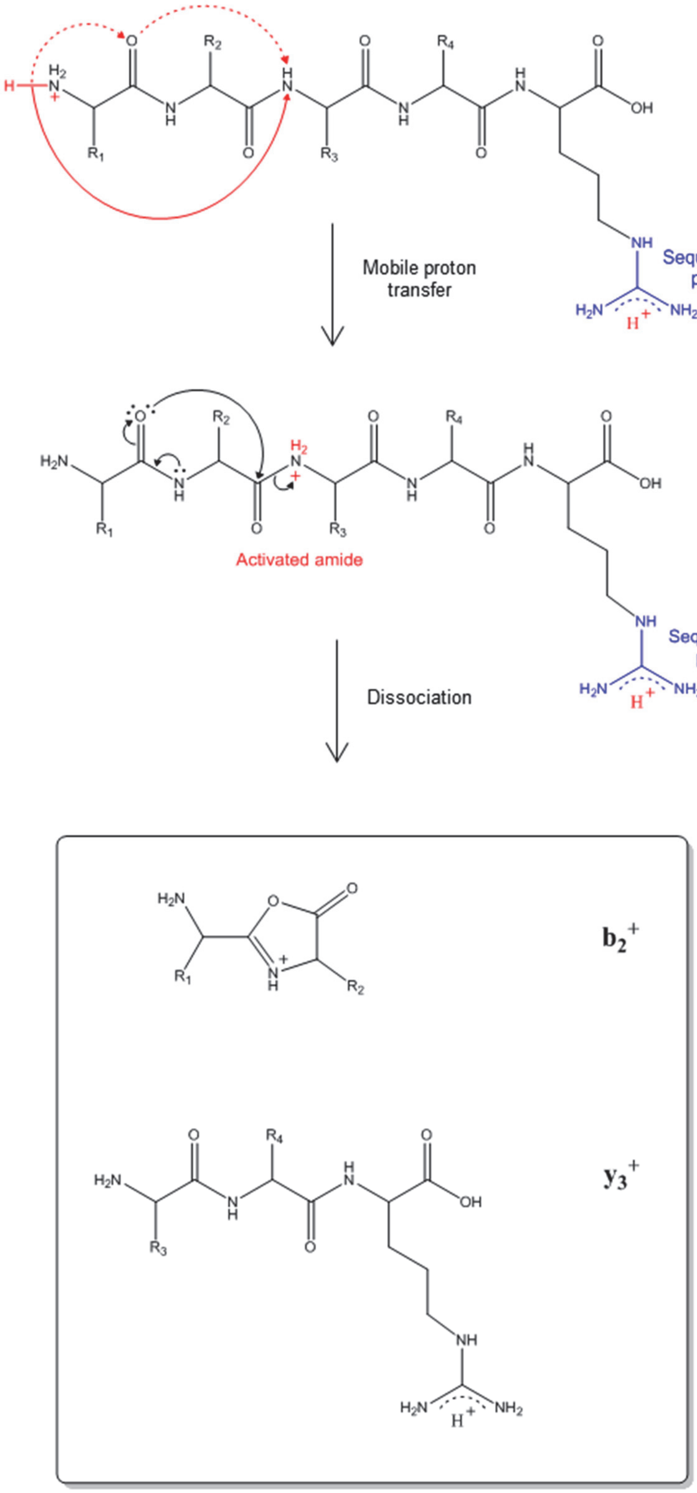


Figure 11: Mobile proton model. Illustration of the dissociation process of a diprotonated tryptic peptide under low-collision energy activation (figure adapted from Paizs et *al.* Mass Spectrometry Review, 2005; 24: 508-48).

Low energy CID is so far the most commonly used activation mode for protein quantification and identification. The presence of PTMs (phosphorylation, glycosylation, sulfonation, etc.) favors CID dissociation pathways orienting the fragmentation towards a neutral loss of the post translational modification, rather that the generation of the specific sequencing ions y- and b- [67].

Electron transfer dissociation (ETD)

ETD is a fragmentation technique introduced in 2004 by Syka et *al.* [68] analogue to electron capture dissociation (ECD). Both techniques are based on a reaction between multi-protonated peptides/proteins with an electron, producing instable radical species that induce their dissociation at the N-C_α bond of the main chain generating principally c and z fragment ions (Figure 12).

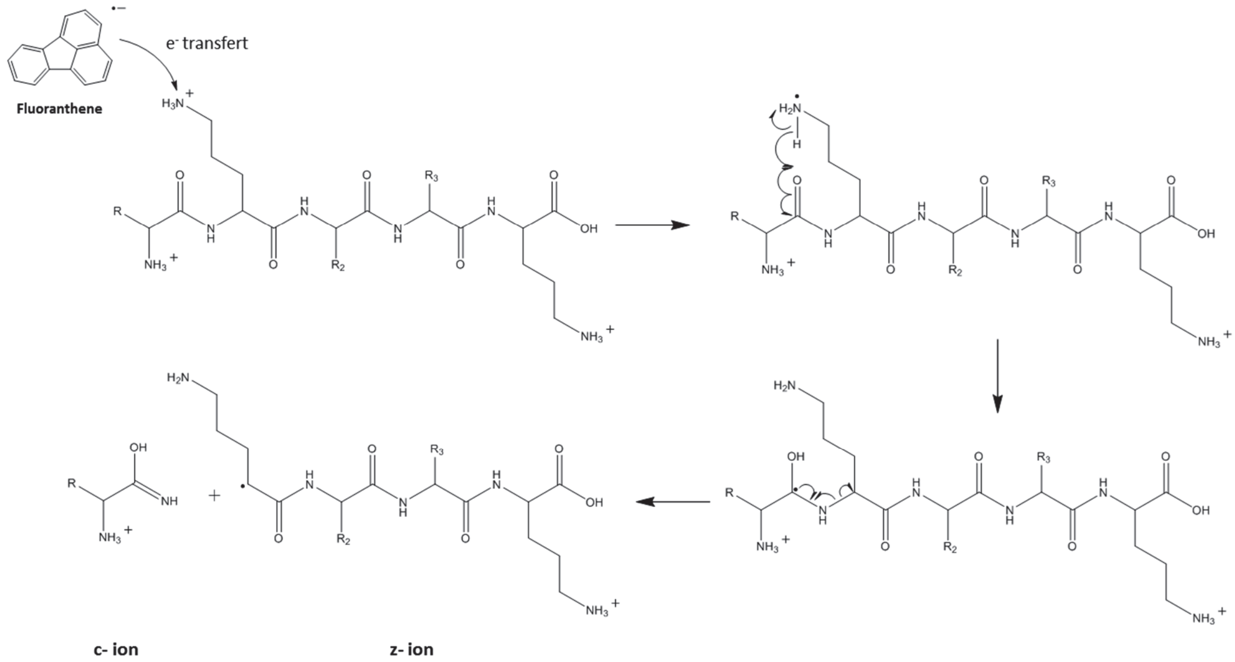


Figure 12: ETD peptide fragmentation pathways for the formation of c- and z- ions (figure adapted from Syka et *al.,* PNSA, 2004; 26: 9528-9533).

In ETD the electron is transferred from a radical anion, usually fluoranthene, produced in a separate chemical ionization source and transferred into a trapping cell where the reduction of multi-protonated peptides/proteins occurs. ETD is not an ergodic activation mode, the excess of energy, which induces the fragmentation is confined to the electron transfer location, unlike to CID where the energy is distributed over all the bonds.

Proteomic strategies

A large majority of proteomic studies rely on mass spectrometric techniques for the identification and the quantification of peptides generated from protein proteolysis. Two main strategies, shotgun and targeted, are mainly used [69].

Bottom-up "shotgun" proteomics [36], in analogy with shotgun genomic sequencing [70], is the most widely used approach for the high-throughput identification of proteins in complex biological samples (Figure 13).

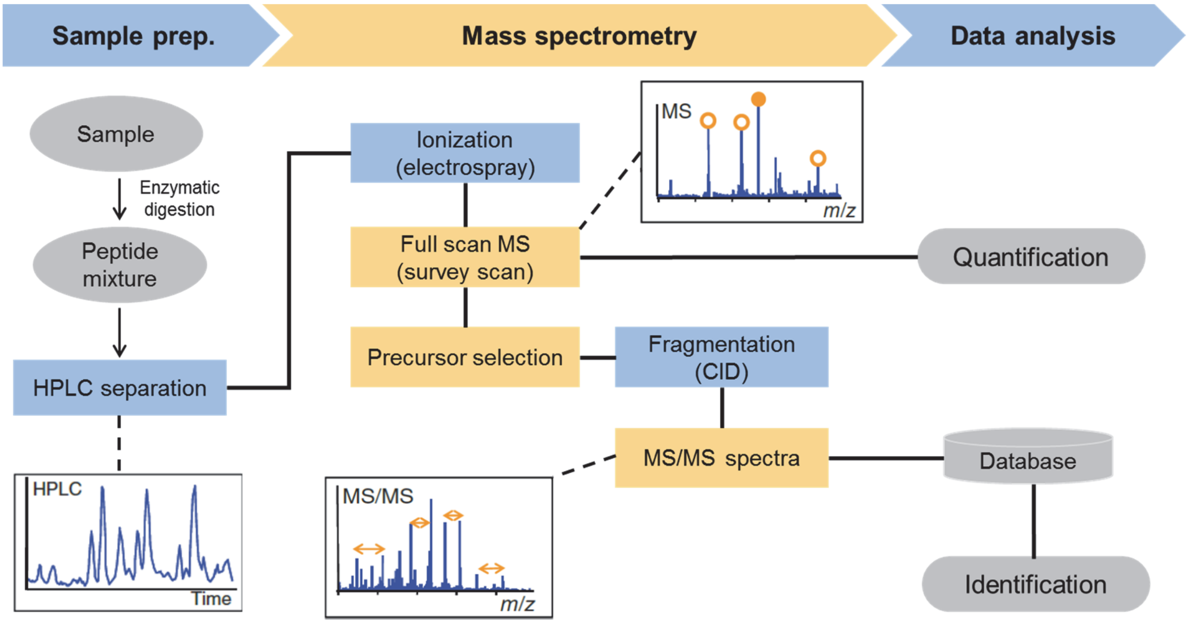


Figure 13: Shotgun proteomic workflow (from Domon et *al.* Nature Biotechnology, 2010; 28: 710–721).

In a standard shotgun experiment, the proteins of a proteome or sub-proteome of interest are digested with trypsin; the resulting peptides are separated according to their hydrophobicity using reverse phase high performance liquid chromatography (HPLC) and finally analyzed by mass spectrometry using the data-dependent acquisition mode (DDA). In DDA experiments, an MS1 survey scan is acquired to determine the most intense ions which will be selected for sequential fragmentation to produce tandem mass spectrometry spectra (MS/MS). The MS1 spectra are also used for quantification and the MS/MS spectra for identification by querying databases using search engines. Using this approach around 4000 proteins were identified in a yeast digest after a one hour gradient using a single dimension chromatographic separation [71]. Shotgun proteomics is the method of choice for proteome characterizations, in spite of some drawbacks for quantitative studies. Indeed, the heuristic nature of the ion selection affects the reproducibility of the method and generally introduces biases towards the most abundant proteins which are amplified due to the complexity and large dynamic range of concentration in biological samples exceeding the peak capacity of the chromatographic separation and the acquisition speed of the mass spectrometer. To overcome these limitations, several strategies such as targeted proteomics were developed to improve the quantification performance.

Targeted proteomics approaches [69] (Figure 14) are hypothesis-driven, in which only a limited set of predefined peptides are measured based on their biological/clinical relevance.

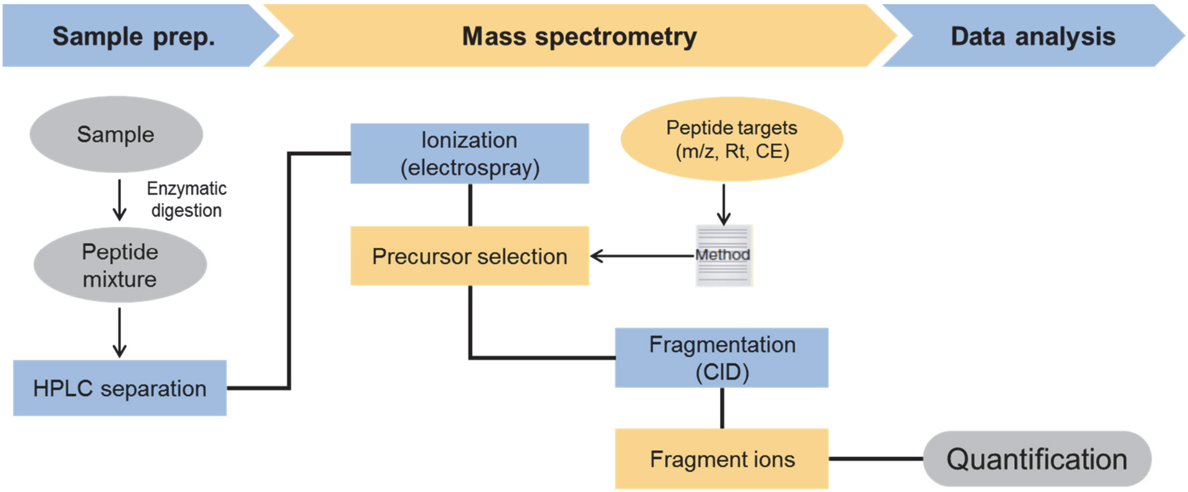


Figure 14: Targeted proteomic workflow (from Domon et *al.,* Nature Biotechnology, 2010; 28: 710–721).

Targeted experiments for protein quantification using surrogate peptides require prior information on the peptide targets. The amino acid sequence has to be unique to the protein of interest (proteotypic) and to present favorable LC-MS(/MS) properties such as the degree of hydrophobicity, flyability in the mass spectrometer and the generated fragmentation patterns. Targeted proteomics experiments are commonly conducted on triple quadrupole mass spectrometers operating in selected reaction monitoring (SRM) mode [72]. The SRM mode was first applied to the analyses of small molecules [73, 74]. More recently it was adapted to the study of peptides for protein quantification in order to achieve the sensitivity and specificity required for the analyses of components at low concentrations in complex biological samples. SRM acquisition is based on the sequential measurement of transitions (pairs of precursor/product ions) (Figure 15).

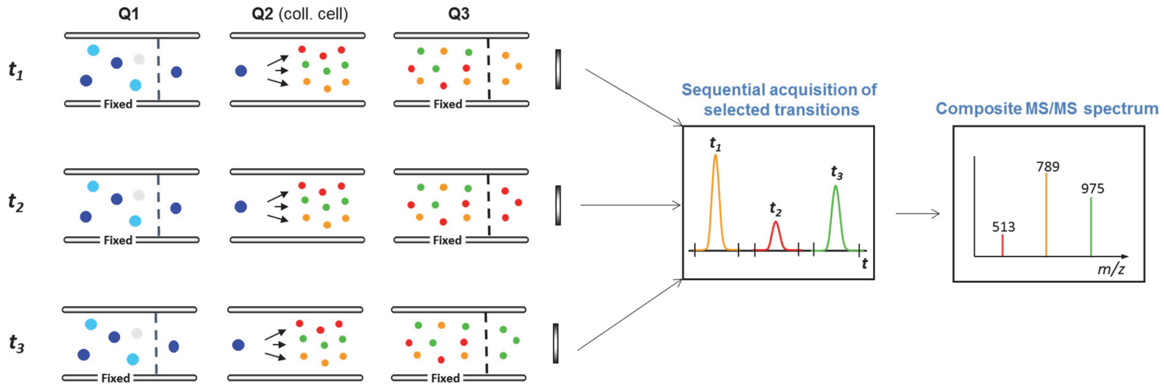


Figure 15: Scheme of a triple quadrupole mass spectrometer working in selected reaction monitoring acquisition mode (adapted from Gallien et *al.,* Journal of Mass Spectrometry, 2011; 3: 298-312).

In the first quadrupole (Q_1), one precursor ion associated to a peptide target is selected, which undergoes fragmentation in the second quadrupole (collision cell) by collision induced dissociation (CID). In the third quadrupole (Q_3), product ions are separated based on their m/z ratio allowing only a specific fragment ion to pass the quadrupole to reach the detector. Peptide quantification is performed using the composite MS/MS spectrum built from the measured transitions. Quantification using SRM on triple quadrupole instruments offers a significant increase in selectivity as compared to shotgun experiments due to the two step mass filtration. However, the low resolution of the quadrupole analyzer does not provide enough selectivity to avoid interferences. The use of the new generation of hybrid mass spectrometers, such as the hybrid quadrupole-orbitrap, which couples a quadrupole analyzer for precursor ion selection with a high resolution mass analyzer for the detection of fragment ions can improve the selectivity of measurement. On quadrupole-orbitrap instruments high resolution targeted MS2 experiments are performed in the parallel reaction monitoring (PRM) [58] acquisition mode (Figure 16).

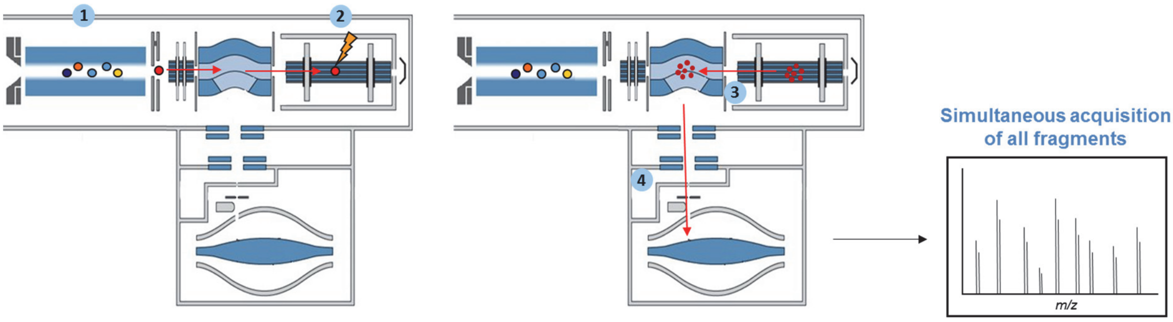


Figure 16: Schema of a quadrupole-orbitrap mass spectrometer operating in parallel reaction monitoring acquisition mode (figure Gallien et *al.,* Molecular Cell Proteomics, 2012; 12:1709-1723).

In PRM, a precursor ion is selected by the quadrupole (1), and transferred to the high energy collisional dissociation (HCD) [75] cell where it is fragmented (2); subsequently all the fragments are transmitted to and accumulated in the C-trap (3) before being transferred into the high resolution orbitrap mass analyzer for detection (4). In PRM experiments, unlike SRM, only a list of precursor ions of interest is required to perform the analysis as all the fragments are recorded simultaneously in the orbitrap analyzer. In this case, peptide quantification is performed in a post-acquisition manner after the extraction of the fragment ions of interest.

Quantitative analyses

The precise quantification of low abundant proteins in complex biological samples remains very challenging, especially the detection of small variations in protein abundance between two physiological states, as required in the context of targeted experiments for the large-scale evaluation of biomarker candidates. Quantitative MS-based proteomic experiments can be performed in two different ways: with relative or absolute quantification.

Relative quantification

In relative quantification (Figure 17), no numerical values for protein concentrations are measured but only the relative variation of protein amount between different samples is recorded. Different strategies have been developed for this approach such as chemical isobaric tagging, metabolic labelling, enzymatic labeling or “label-free” [76].

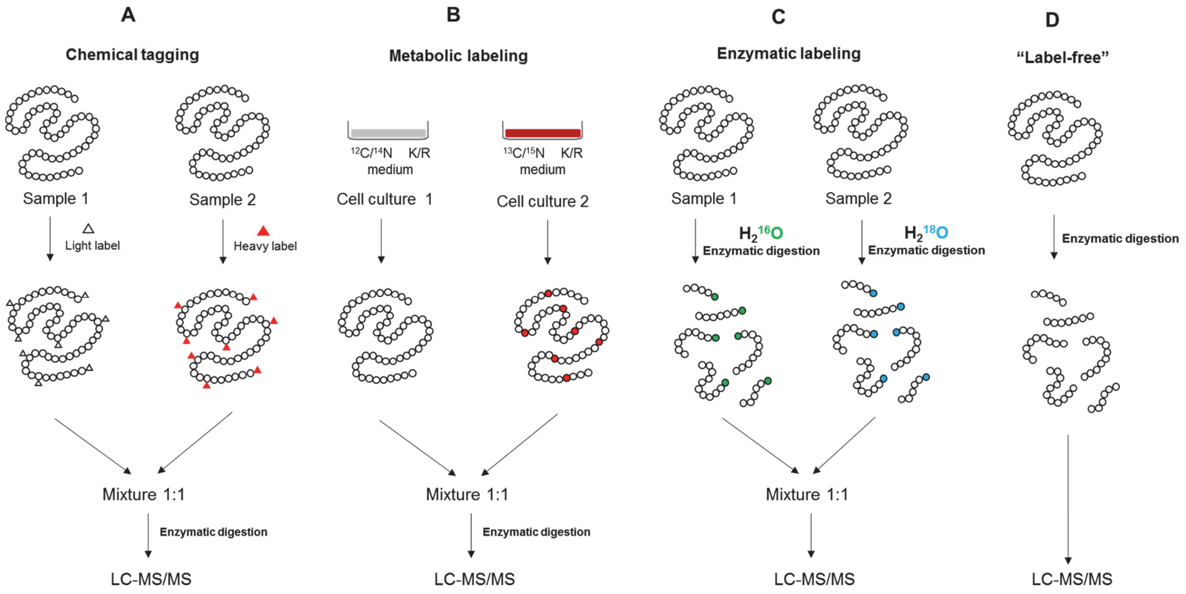


Figure 17: Illustration of proteomic workflows for relative protein quantification. Chemical tagging (A), metabolic labeling (B), enzymatic labeling (C), and “label-free” (D) (figure from Hawkridge, Quantitative Proteomics, 2014; chapter 1: 1-25)

Chemical tagging strategies (Figure 17A) comprise the pooling of different protein samples, each one derivatized with a different isotope tag having the same chemical structure but different mass induced by the incorporation of heavy stable isotopes ^{2}H, ^{13}C or ^{15}N. Based on this approach several strategies were developed such as isotope-coded affinity tag (ICAT) [77], tandem mass tags (TMT) [78] or isobaric tags for relative and accurate quantification (iTRAQ) [79] having multiplexing capabilities of 2, 6, and 8, respectively. The multiplexing degree of a technique is the maximum number of samples on which the relative quantification can be performed. In ICAT the chemical modification occurs on the cysteine residues of proteins using two different tags containing a reactive thiol group, a biotin group and a linker containing eight hydrogens ^{1}H or eight hydrogens ^{2}H introducing a 8 Da mass shift between the two tags which can be differentiated by mass spectrometry. In TMT and iTRAQ strategies, the chemical tag is composed of a unique mass reporter (different combinations of ^{13}C and ^{15}N labelling), a cleavable linker plays the role of a mass balancer to have a constant total mass between each tag and a reactive chemical group able to react with lysine and N-terminal parts of peptides. During the MS/MS experiment, the linker is fragmented, releasing the reporter indicating the relative abundance of the initial peptide. These approaches ensure a good reproducibility because all the analytes are measured simultaneously but are subject to an increased number of sample preparation steps introducing variability which can affect the accuracy of the measurements.

The metabolic labeling (Figure 17B) for relative quantification is based on the growth of a living organism in a medium enriched with stable isotopes. In this context, during all the cellular processes, the stable isotopes will be incorporated by the cells. For example, the stable isotope labeling with amino acids in cell culture (SILAC) [80] strategy introduced in 2002 uses this workflow. Two cell cultures are grown in different media, one containing unlabeled lysine or arginine and the other one containing isotopically labeled lysine or arginine. The two samples containing labeled and unlabeled proteins are pooled, enzymatically digested to release labeled and unlabeled peptides and analyzed by mass spectrometry. The intensity ratio between labeled and unlabeled peptides reflects the relative abundances of the proteins in the initial samples. SILAC is an *in-vitro* approach and hence cannot be easily applied to clinical proteomic samples such as plasma, urine or human tissues.

Enzymatic labeling (Figure 17C) is another stable isotope labeling method. In this technique, the peptide labeling is realized during the proteolysis of proteins in ^{18}O water [81]. Indeed, during the catalytic mechanism of hydrolysis of proteins by serine proteases one oxygen atom from a water molecule is incorporated into the C-terminal part of the released peptides. A second atom of oxygen can be incorporated via a reaction of carboxyl oxygen exchange catalyzed by some of the enzymes. Thus, an enzymatic digestion of proteins in ^{18}O water can generate peptides with 2 or 4 Da mass shift (trypsin yields mainly + 4Da) in comparison to a digestion performed in regular ^{16}O water. The main challenge of ^{18}O labeling is to achieve an efficient and quantitative incorporation of ^{18}O during proteolysis; several methods have been developed to specifically incorporate one or two ^{18}O atoms [82]. This approach is generally applicable to biological samples for relative quantification, unlike SILAC. The main limitation of ^{18}O labeling is the small mass shift between labeled and unlabeled peptides that can induce an overlap of their isotopic distributions making quantification more challenging [83].

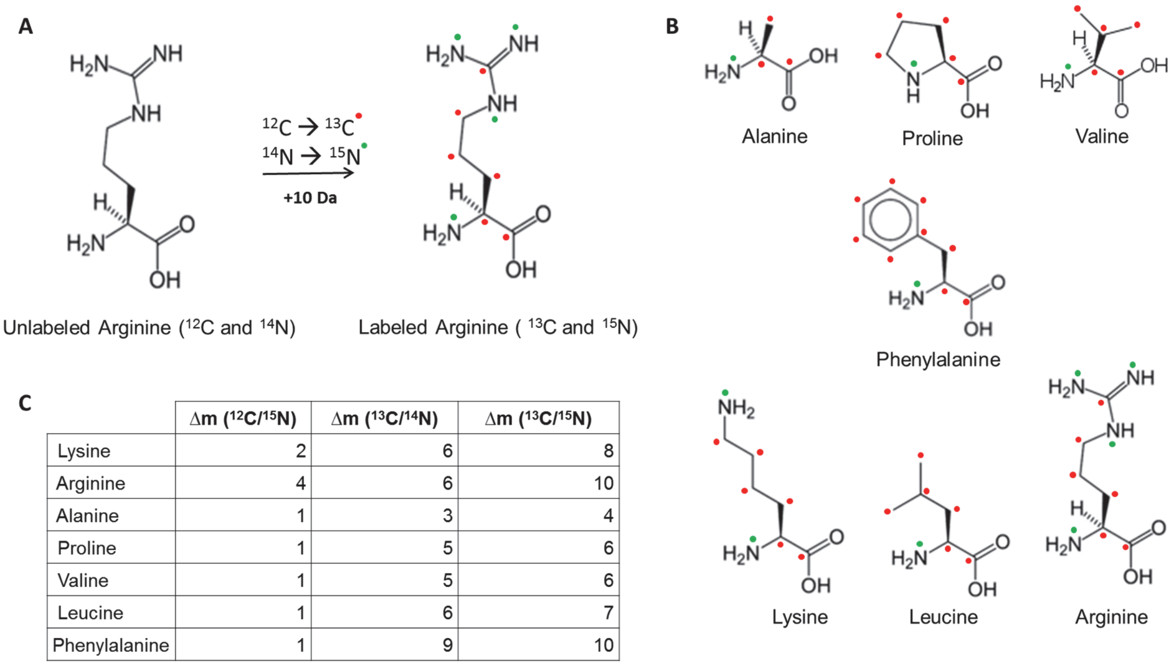
Unlike other relative quantification strategies, label-free approaches [84] (Figure 17D) do not use stable isotopes. Only an enzymatic digestion is required before the analysis by LC-MS/MS in DDA mode. The relative peptide quantification is performed post-acquisition using spectral counting or ion intensity measurements. Spectral counting [85] is based on the hypothesis that, after a database search, the most abundant proteins have a larger number of identified peptides, a higher protein coverage and an increased number of MS/MS assignments than low abundant proteins. In ion intensity measurements [86], the quantification is based on the intrinsic relationship between ion intensity and concentration which are linearly correlated.

Label-free quantification is a straightforward approach which allows to quantify an important number of samples but requires highly reproducible instrumental conditions [87].

Absolute quantification

In absolute quantification, in contrast to relative quantification, the result of the protein abundance determination process is a numerical value (*e.g.,* pmol/µL) which allows an independent treatment of samples. If the determined value is very close to the real one, the quantification will be qualified as accurate. Absolute and accurate quantifications are required in a plurality of scientific fields, for example in the context of protein diagnostic biomarkers in biological samples for clinical tests with the establishment of protein concentration thresholds. Absolute quantification strategies based on mass spectrometry can be performed in two different ways, using stable isotope labels as internal standards or without internal standards via a label-free approach [88].

The use of stable isotope dilution combined with SRM acquisition is currently the gold standard strategy for accurate protein quantification. In this approach, known amounts of isotope-labeled peptides, corresponding to the proteins of interest, are added to the biological/clinical sample. The internal standards have the same physicochemical properties as the targets in order to have identical separation behavior during all the sample processing, *i.e.,* the chromatographic separation and the mass spectrometry analysis. The isotope labeled internal standards have the same amino acid sequences as the endogenous targets but with different molecular masses due to the incorporation of heavy stable isotopes (^{13}C and/or ^{15}N) in certain amino acids (Figure 18). This incorporation of stable isotopes generates a mass shift between the endogenous target and the internal standard making them differentiable in mass. Several amino acids can be isotopically modified but usually the labeling is performed on the C-terminal amino acids, lysine or arginine, for peptides generated by trypsin, for synthesis reasons and for peptide fragmentation considerations.



	Δm (^{12}C/^{15}N)	**Δm (^{13}C/^{14}N)**	**Δm (^{13}C/^{15}N)**
Lysine	2	6	8
Arginine	4	6	10
Alanine	1	3	4
Proline	1	5	6
Valine	1	5	6
Leucine	1	6	7
Phenylalanine	1	9	10

Figure 18: Stable isotope-labeled amino acids. Panel A presents the full labelling of arginine amino acid with ^{13}C (red dot) and ^{15}N (green dot), B displays the structure of six common amino acids and C shows the mass shift induced by three stable isotope labels ^{15}N, ^{13}C and ^{13}C/^{15}N.

Several types of stable isotope-labeled (SIL) standards have been proposed for accurate quantification, including: calibrated isotopically-labeled peptides (CIP) also called AQUA, extended peptide concatamers (QconCAT), or full-length proteins (PSAQ) [89] (Figure 19).

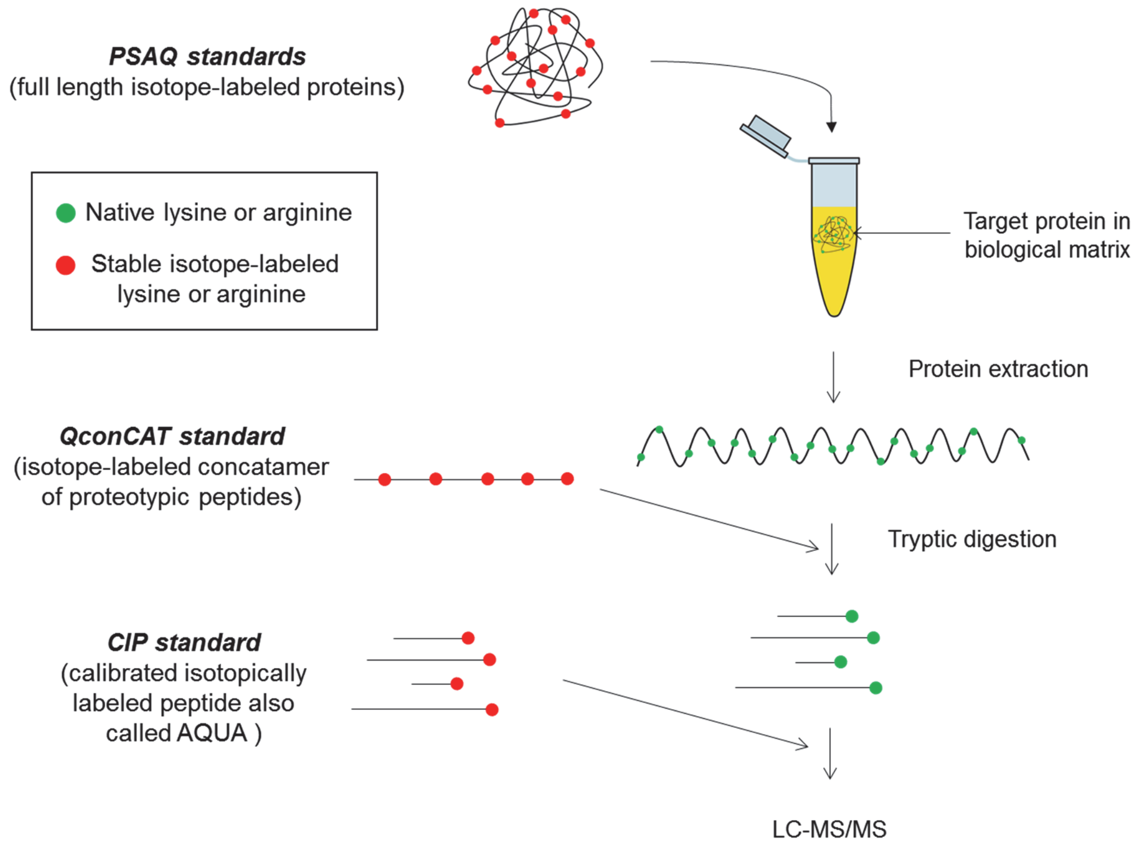

PSAQ standards
(full length isotope-labeled proteins)
Native lysine or arginine
Stable isotope-labeled lysine or arginine
Target protein in biological matrix
Protein extraction
QconCAT standard
(isotope-labeled concatamer of proteotypic peptides)
Tryptic digestion
CIP standard
(calibrated isotopically labeled peptide also called AQUA)
LC-MS/MS

Figure 19: Stable isotope labeled strategies for accurate protein quantification using targeted LC-MS/MS experiments. Three standards are presented, PSAQ (isotope labeled protein) added directly to the biological sample, QconCAT (isotope labeled concatamer of proteotypic peptides) incorporated just before the enzymatic digestion, and CIP (calibrated isotopically labeled proteotypic peptides) supplemented in the biological digest before the LC-MS/MS analyses (figure adapted from Brun et *al*., Journal of Proteomics, 2009; 5: 740-9).

CIP [90] are chemically synthesized peptides containing amino acids isotopically-labeled with the heavy stable isotopes ^{13}C and ^{15}N. For an accurate quantification, synthetic CIP require a high degree of purity and their concentration needs to be determined accurately. The major advantage of CIPs is their straightforward utilization because they generally are added only before LC-MS analyses. This strategy suffers from several limitations that can affect protein quantification. First, the stability of the standards over time can affect their nominal concentration due to unspecific adsorption interactions with the container or degradation. Second, certain synthetic peptides are difficult to produce and purify especially for large and very hydrophobic peptides. QconCAT concatemers [91] are labeled polypeptides resulting from the fusion of up to ten proteotypic peptides coming from different proteins produced by recombinant expression of a synthetic gene, with isotopically-labeled amino acids. This approach allows the multiplexing of several protein targets. QconCAT internal standards are added before the enzymatic digestion, thus for an accurate quantification a complete proteolysis of the standard is required to keep the equimolarity between the released isotopically-labeled prototypic peptides. Moreover, this stoichiometry does not allow the simultaneous quantification of proteins having a large concentration difference. The PSAQ standards [92] are a full-length versions of the proteins of interest containing heavy lysine and arginine residues. PSAQ standards are directly spiked into the biological samples and will undergo the same treatment as the endogenous proteins, with the same loss during the sample preparation and the same digestion efficiency. An important limitation of PSAQ standards is that they do not truly reflect the endogenous proteins especially if these are subject to various isoforms or PTMs. The use of stable isotope-labeled internal standards confers in addition to accurate quantification a high degree of reliability, precision, sensitivity and reproducibility.

The different information and methodologies described in this introduction part were used to develop alternative proteomic methods to overcome the weakness of the standard proteomic workflows. First, a quantification strategy using concatenated polypeptides was developed to improve accuracy, second an assessment of the advantages of alternative enzymes was conducted and to finish the developed strategies were combined to perform an accurate quantification of NSCLC biomarkers in clinical plasma samples.

Chapter I: A quantification strategy using concatenated polypeptides

This chapter describes experiments published in the article *Protein quantification using a cleavable reporter peptide*, E. Duriez, S. Trévisiol, B. Domon, Journal of Proteome Research, 2015; 2: 728-37 (Annex 3).

1. Introduction

In quantitative assays involving the use of calibrated isotope-labeled peptide (CIP) standards, the reliability and the accuracy of the experiments is dependent on their quality. Indeed, for accurate quantification high isotope incorporation, high chemical purity, and an accurate concentration of the standards are required [93]. Calibrated isotope-labeled peptides are often employed in quantitative experiments due to their ease of use [94-98] but they can also bias the quantification due to uncontrolled losses of material before their addition into biological samples [99] which affects accuracy. Different factors may affect the nominal amount of calibrated isotope-labeled peptides added to samples between their synthesis and their utilization, *e.g.* incomplete solubilization, non-specific adsorption on hydrophobic surfaces or aggregation [100, 101]. These effects can be accentuated according to the buffer formulation, the material of the storage vials, or the storage conditions of the calibrated standards [102]. Protein quantification usually relies on the nominal amount of calibrated isotope-labeled peptides provided by the manufacturer, determined by photometric methods or by quantitative amino acid analysis (AAA) [103].

For accurate quantification, the concentration of the internal standard should be recalibrated before each experiment. Different strategies using isotopically-labeled signature peptide (ISP) standards based on the equimolar generation of products after proteolysis were developed [100, 104, 105]. Synthetic concatenated polypeptides (CPP) containing a trypsin cleavage site can be used to determine isotopically-labeled signature peptide standard amounts just before quantitative experiments. In this context, we developed a new quantification strategy using concatenated polypeptide standards.

2. The concatenated polypeptide concept

The principle of the developed method relies on the concatenation of two peptides: an isotopically-labeled signature peptide (ISP) and a cleavable reporter peptide (RP) (Figure 20).

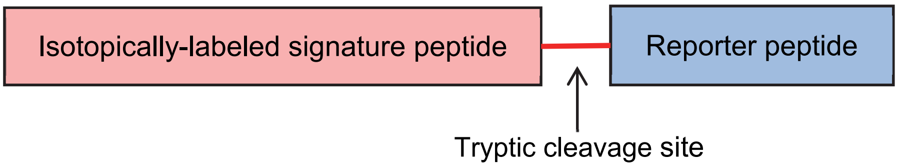


Figure 20: Structure of a concatenated polypeptide standard.

The quantification using a concatenated polypeptide as an internal standard is based on the equimolar amount (stoichiometry 1:1) of its two constituents: the ISP and the cleavable RP. During the tryptic digestion, the concatenated polypeptide releases the signature peptide, which is the surrogate of the protein of interest, and the cleavable reporter in an equimolar amount. The accurate quantification of the isotopically-labeled signature peptide as internal standard is performed via the quantification of the cleavable reporter using a calibration mixture containing reporter isotopologues in increasing concentrations (Figure 21). The isotopologue peptides have the same amino acid sequence, and thus the same physicochemical properties (hydrophobicity, ionization factor, fragmentation behaviors), but differ in their molecular mass due to the incorporation of heavy stable isotope amino acid residues at various positions within the sequence.

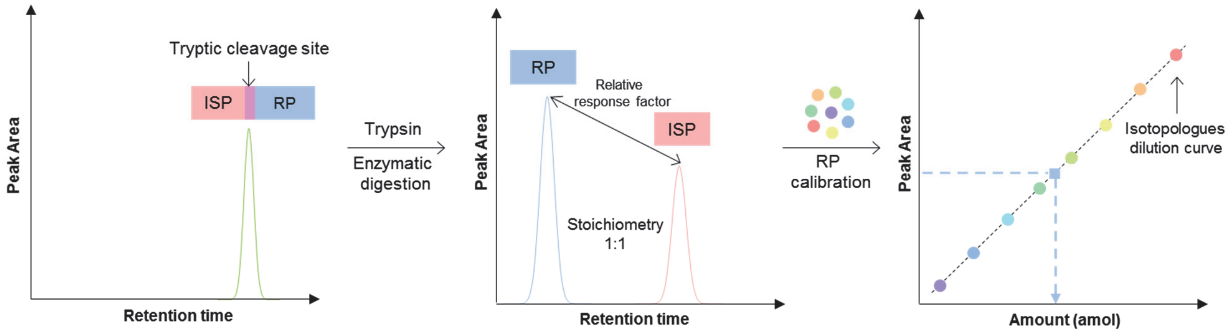


Figure 21: Description of the concatenated polypeptide approach (figure adapted from Duriez et *al.* Journal of Proteome Research, 2015; 2: 728-37).

The amino acid sequence of the reporter is common to all polypeptide standards, which enables the systematic calibration or recalibration of any of these sequence-specific signature peptides after the determination of the relative response factor (RRF) of the polypeptide digest. The RRF is specific to each concatenated polypeptide digest and defined as the ratio of intensities of the isotopically-labeled signature peptide with the cleavable reporter which reflects the difference of ionization efficiency of the two partners (Figure 22).

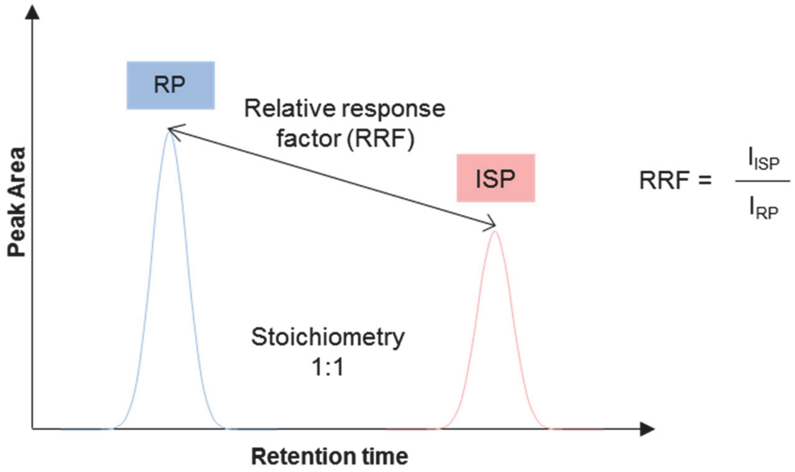

RP
Relative response factor (RRF)
ISP
RRF = $\frac{I_{ISP}}{I_{RP}}$
Peak Area
Stoichiometry 1:1
Retention time

Figure 22: Illustration of the relative response factor (RRF) defined as the ratio of intensities between the isotopically-labeled signature peptide (ISP) and the cleavable reporter peptide (RP) released in equimolar amount after the tryptic digestion of the concatenated polypeptide.

The first generation of concatenated polypeptide standards has been presented during the annual conference of the American Society for Mass Spectrometry (ASMS) in 2011 as a novel approach for precise protein quantification in complex biological samples [106]. Then this strategy was communicated the same year in the context of the annual conference of the Human Proteome Organization (HUPO) for the quantification of lung cancer biomarker candidates in clinical plasma samples [107]. In this first generation of the standards, the cleavable reporter was a 7 amino acid peptide (LVALVR) fused to the N-terminal part of the signature peptide. With this structure, several issues were reported. First, depending on the signature peptide used, variable digestion efficiencies were observed, from 0% to 100%. The poor digestion efficiencies were attributed to the amino acid sequence of the cleavable reporter and the N-terminal position in the standard. Second, the reporter peptide (LVALVR) was found to be too short for the design of the isotopologue calibration mixture. In this context, to overcome these limitations a new amino acid sequence for the reporter was designed.

3. Design of the reporter amino acid sequence

The selection of the amino acid sequence of the reporter peptide was based on different criteria including LC-MS detectability (amino acid composition, hydrophobicity), trypsin digestibility specificity (the presence of acidic amino acids close to the cleavage site can affect proteolysis efficiency), number of amino acids (synthesis criteria) and cost of synthesis (availability of isotope-labeled amino acids, L, V and A were prioritized). Moreover, in order to be able to be directly calibrated in biological samples, the amino acid sequence of the reporter cannot be present in the UniProt KB database [Version 2011_10]**.** With these considerations in mind, the selection and the evaluation of the best amino acid sequences for the reporter peptide were performed in three steps.

Selection of reporter peptide candidates

The first step in the design of the reporter peptide was the selection of a broad set of candidates generated in-silico based on their amino acid sequence before experimental confirmation. As previously mentioned, length and amino acid composition of the reporter are crucial for a high digestion efficiency of the concatenated polypeptide and for a reliable quantification of the isotopically-labeled signature peptide. The new reporter was defined as a unique, non-occurring in Swissprot database (all species), eight amino acid tryptic peptide created by concatenating two tetrads of amino acids XXXX and XXXK/R observed in the human proteome (where X represents any amino acid except lysine or arginine). Using an in-house developed software, all the possible combinations for the two tetrads were listed. Only tetrads observed in the human proteome were retained i*.e.,* 104669 possibilities for XXXX and 11659 for XXXK/R. The amino acid sequence resulting from the fusion of the two tetrads has to be unique through various species. So the different tetrads should not be observed too many times to ensure the uniqueness of the concatenated product. To satisfy the observability criteria, the choice was made to select for the two tetrads only the amino acid sequences which are observed close to the mean of observations. On average, the tetramers XXXX and XXXK/R are observed 67 times (523 combinations) and 72 times (61 combinations), respectively. At this point, 31903 amino acid sequences were possible for the reporter and 20488 after the rejection of those containing methionine or cysteine residues. A BLAST (Basic Local Alignment Search Tool) of these sequences was performed against *Swissprot* (all species) to exclude non-unique tryptic sequences. In the final selection step of the reporter sequences, synthesis and liquid chromatography considerations were taken into account. Only peptides having a hydrophobicity factor (Krokhin [108, 109]) between 10 and 15 were considered. Moreover, peptides containing several alanine, leucine and valine were preferred. From this selection process, a set of 280 eight amino acid peptides were synthetized and evaluated experimentally by LC-MS.

The 280 reporter candidates were separated in 20 sub-sets of 14 peptides and analyzed by LC-MS on the Q-Exactive mass spectrometer platform to determine peptides with the best MS response factor (flyability) based on MS1 signals and chromatographic profiles. At this point 19 peptides were retained (Table 1).

Table 1: List of 19 reporter candidates selected for the evaluation of concatenated polypeptide proteolysis efficiency (table adapted from Duriez et *al.* Journal of Proteome Research, 2015; 2: 728-37).

AALFAATK	AALQAAFK	AANFAAFK	AAQFAAFK	AAYFAAVK
AALHAAFK	AALQAAWK	AANLAAFK	AAQLAALK	VVAPLVAK
AALPAAFK	AALWAATK	AANLAALK	AAVLAAFK	VVAPVVAK
AALPAAWK	AALYAAYK	AAPFAAFK	AAYFAAFK	

Evaluation of reporter sequences on concatenated polypeptides proteolysis efficiency

The second step in the design of the new reporter consisted in the assessment of the proteolysis efficiency of the concatenated polypeptide when the reporter is concatenated to the signature peptide. Three isotopically-labeled signature peptides, SFFSFLGEAFDGAR (ISP1), ELDESLQVAER (ISP2), and ASSIIDELFQDR (ISP3) were selected to be fused with the 19 reporters either at their C-terminus or N-terminus. The selection of the three ISPs was based on their amino acid sequences with the presence of aspartic and glutamic acid residues well known to increase the missed-cleavage rate when located close to the tryptic cleavage site [110]. Polypeptides were digested by trypsin individually. Digestion efficiency of the 107 concatenated polypeptides was measured by LC-MS (SRM mode) on a triple quadrupole mass spectrometer by monitoring the amount of undigested standards after trypsinization (Table 2). Regarding the position of the cleavable reporter within the concatenated polypeptide sequence, the fusion of the reporter at the C-terminus of the signature peptide was optimal. Higher digestion efficiency was observed when the reporter was located in front of the standard; particularly peptide AALPAAFK showed a proteolysis efficiency >99%, regardless of the signature peptide used. According to these results, the AALPAAFK amino acid sequence was selected as reporter and its position was fixed at the C-terminal part of the concatenated polypeptide.

Table 2: Digestion efficiency of 107 cleavable polypeptides. Green color highlights digestion efficiency above 99% and grey color corresponds to the polypeptides which presented synthesis difficulties.

RP	RP---ISP1	RP---ISP2	RP---ISP3	ISP1---RP	ISP2---RP	ISP---RP
AALFAATK	99.1					99.7
AALHAAFK	96.3		99.8	0.0	99.8	93.3
AALPAAFK	96.8	57.5	96.4	100.0	99.0	99.9
AALPAAWK	99.3		100.0	99.8	96.3	97.0
AALQAAFK	99.8	56.4	99.2	99.4	51.4	99.5
AALQAAWK	46.0	90.8	95.4	0.0	82.6	82.6
AALWAATK	99.2	87.3	100.0	0.0	78.6	99.1
AALYAAYK	88.5	92.5	100.0	9.2	99.3	
AANFAAFK	99.7	2.6	97.5	98.8	99.2	96.0
AANLAAFK	94.1	84.3	86.0	100.0	84.1	92.6
AANLAALK	60.7	61.6	94.1	100.0	68.0	97.4
AAPFAAFK	92.8	92.3	99.4	100.0	79.4	99.7
AAQFAAFK	20.0	55.3	97.0	87.4	3.0	99.4
AAQLAALK	90.3	71.0	92.0	98.9	93.4	99.9
AAVLAAFK	89.1	47.8	89.2	0.00	91.0	95.6
AAYFAAFK	91.6	28.5	100.0	64.3	99.9	99.9
AAYFAAVK	92.9	42.4	92.2	0.00	98.9	95.2
VVAPLVAK	97.1	73.6	81.8	100.0	61.4	95.7
VVAPVVAK	95.4	69.7	98.0	99.5	83.8	83.6

Evaluation of the concatenated polypeptides digestion efficiency

In order to confirm the choice of AALPAAFK as the reporter peptide, the digestion efficiency of 46 concatenated polypeptide sequences was tested. As before, the proteolysis efficiency of the different polypeptides was measured by LC-MS (SRM mode) on a triple quadrupole instrument (Table 3). The efficiency of the digestion observed for the 43 polypeptides, containing the AALPAAFK reporter sequence, was on average 96% (CV 8.6%). The new reporter, unlike the initial LVALVR, results in an efficient proteolysis of concatenated polypeptide standards regardless of the signature peptide sequence to which it is concatenated.

Table 3: Digestion efficiency of 46 concatenated polypeptides which contain AALPAAFK as cleavable reporter (table adapted from Duriez et *al.* Journal of Proteome Research, 2015; 2: 728-37).

CPP	Sequence	Proteolysis efficiency (%)
ISP1- RP	SFFSFLGEAFDGAR-AALPAAFK	99
ISP2- RP	ELDESLQVAER-AALPAAFK	99
ISP3- RP	ASSIIDELFQDR-AALPAAFK	99
ISP4- RP	DWVSVVTPAR-AALPAAFK	100
ISP5- RP	SGSVIDQSR-AALPAAFK	95
ISP6- RP	DGAGDVAFVK-AALPAAFK	97
ISP7- RP	SASDLTWDNLK-AALPAAFK	99
ISP8-RP	LADGGATNQGR-AALPAAFK	100
ISP9-RP	SDLAVPSELALLK-AALPAAFK	99
ISP10-RP	TVGSDTFYSFK-AALPAAFK	99
ISP11-RP	YFIDFVAR-AALPAAFK	100
ISP12-RP	YDLLDLTR-AALPAAFK	99
ISP13-RP	SDVFEAWR-AALPAAFK	99
ISP14-RP	HFTYLR-AALPAAFK	88
ISP15-RP	AWTDVLPWK-AALPAAFK	97
ISP16-RP	DWVSVVTPAR-AALPAAFK	72
ISP17-RP	DSTIQVVENGESSQGR-AALPAAFK	96
ISP18-RP	FAGVFHVEK-AALPAAFK	99
ISP19-RP	YGFIEGHVVIPR-AALPAAFK	100
ISP20-RP	LVDQNIFSFYLSR-AALPAAFK	97
ISP21-RP	STITLDGGVLVHVQK-AALPAAFK	100
ISP22-RP	TAFYLAEFFVNEAR-AALPAAFK	100
ISP23-RP	YPVYGVQWHPEK-AALPAAFK	100
ISP24-RP	HEVTGWVLVSPLSK-AALPAAFK	100
ISP25-RP	LKPEDITQIQPQQLVLR-AALPAAFK	99
ISP26-RP	SDLVNEEATGQFR-AALPAAFK	100
ISP27-RP	CETQNPVSRA-AALPAAFK	99
ISP28-RP	GEHGFIGCR-AALPAAFK	99
ISP29-RP	HYGYNSYSVSNSEK-AALPAAFK	99
ISP30-RP	LPASFDAR-AALPAAFK	93
ISP31-RP	VISSIEQK-AALPAAFK	100
ISP32-RP	YDEEFASQK-AALPAAFK	55
ISP33-RP	SGSVIDQSR-AALPAAFK	99
ISP34-RP	ALSIGFETCR-AALPAAFK	95
ISP35 RP	VSTLPAITLK-AALPAAFK	99
ISP36-RP	YSQAVPAVTEGPIPEVLK-AALPAAFK	100
ISP37-RP	LLLTSAPSLATSPAFR-AALPAAFK	99
ISP38-RP	YDLLDLTR-AALPAAFK	100
ISP39-RP	EVGVGFATR-AALPAAFK	98
ISP40-RP	HLDSVLQQLQTEVYR-AALPAAFK	98
ISP41-RP	LTIESTPFNVAEGK-AALPAAFK	99
ISP42-RP	QIGDALPVSCTISASR-AALPAAFK	90
ISP43-RP	AVTELNEPLSNEDR-AALPAAFK	79
ISP44-RP	NLLSVAYK-AALPAAFK	99
ISP45-RP	SDVFEAWR-AALPAAFK	100
ISP46-RP	APAVAEENPK-AALPAAFK	86

4. Design of the reporter isotopologue calibration mixture

A reliable calibration of the cleavable reporter requires the use of a calibration mixture containing at least four different isotopologues of precisely quantified increasing concentration to establish a dilution curve. For concomitant analysis the mass difference between isotopologues has to be sufficient to avoid overlap of isotopes from one variant to the next one, considering a doubly charge precursor and a mass selection (triple quadrupole, 1 m/z).

Selection of AALPAAFK isotopologues

The reporter peptide is an eight amino acid peptide composed of five distinct amino acids, alanine (A), proline (P), leucine (L), lysine (K) and phenylalanine (F). The heavy stable isotope labelling with ^{13}C and ^{15}N of these five amino acids induces a mass shift of 4 Da, 6 Da, 7 Da, 8 Da and 10 Da, respectively. These amino acid sequence types of isotope labeling, allow for 256 different combinations of isotopologues variants. All the possibilities were listed and grouped by molecular mass. One representative of each isotopologue group was selected for synthesis. The choice was based on the position of the heavy stable isotope labeled amino acids within the sequence. In CID for doubly charged tryptic peptides without internal basic amino acid residues fragmentation is dominated by y-ions. For this reason isotopologue peptides with the maximum number of stable isotope labeled amino acids close to the C-terminal end were retained. These criteria allowed to select 38 isotopologues, which cover a mass range of 47 Da (Table 4).

Table 4: List of 38 synthetic stable isotope labeled peptide variants of AALPAAFK with various combinations of ^{13}C and ^{15}N labeled amino acids (labeling in red).

Sequence	Molecular mass	m/z (+2)	∆m
AALPAAFK	787.459	394.737	0
AALPAAFK	791.466	396.740	4
AALPAAFK	793.473	397.744	6
AALPAAFK	794.476	398.245	7
AALPAAFK	795.473	398.744	8
AALPAAFK	797.486	399.751	10
AALPAAFK	798.483	400.249	11
AALPAAFK	799.481	400.748	12
AALPAAFK	800.490	401.252	13
AALPAAFK	801.494	401.754	14
AALPAAFK	802.491	402.253	15
AALPAAFK	803.488	402.751	16
AALPAAFK	804.497	403.256	17
AALPAAFK	805.501	403.758	18
AALPAAFK	806.498	404.256	19
AALPAAFK	807.507	404.761	20
AALPAAFK	808.511	405.263	21
AALPAAFK	809.508	405.761	22
AALPAAFK	810.505	406.260	23
AALPAAFK	811.514	406.765	24
AALPAAFK	812.518	407.266	25
AALPAAFK	813.515	407.765	26
AALPAAFK	814.525	408.270	27
AALPAAFK	815.522	408.768	28
AALPAAFK	816.525	409.270	29
AALPAAFK	817.522	409.768	30
AALPAAFK	818.532	410.273	31
AALPAAFK	819.529	410.772	32
AALPAAFK	820.532	411.273	33
AALPAAFK	821.529	411.772	34
AALPAAFK	822.539	412.277	35
AALPAAFK	823.536	412.775	36
AALPAAFK	824.539	413.277	37
AALPAAFK	826.546	414.28	39
AALPAAFK	827.543	414.779	40
AALPAAFK	828.546	415.280	41
AALPAAFK	830.553	416.284	43
AALPAAFK	834.560	418.287	47

Co-elution of reporter peptide isotopologues

Peptide isotopologues have the same amino acid sequence but differ by their number of neutrons. Thus, they have the same physicochemical properties namely: hydrophobicity, ionization factor and fragmentation behavior. It has been reported that in the case of deuterium labeling, shifts in retention time between the different isotopologues are observed. This chromatographic deuterium shift was observed for small organic molecules [111] and also for peptides [112], for example in the relative quantification approach ICAT which involves a deuterium labeling. To overcome the retention time issue, cICAT [113] labels have been developed using ^{13}C stable isotopes, instead of deuterium, for which no retention time shift was observed. Regarding the reporter peptide AALPAAFK, the fully labeled isotopologue corresponds to the addition of 47 neutrons. Theoretically, labeling with the stable isotopes ^{13}C and ^{15}N may not induce a retention time shift, but due to the mass increase, modifications of hydrophobic interactions can potentially occur. To ensure the co-elution of all the reporter isotopologues, they were pooled and analyzed by LC-MS on the Q-Exactive mass spectrometer. The chromatogram in Figure 23 shows the co-elution of all the isotopologues, regardless of the mass.

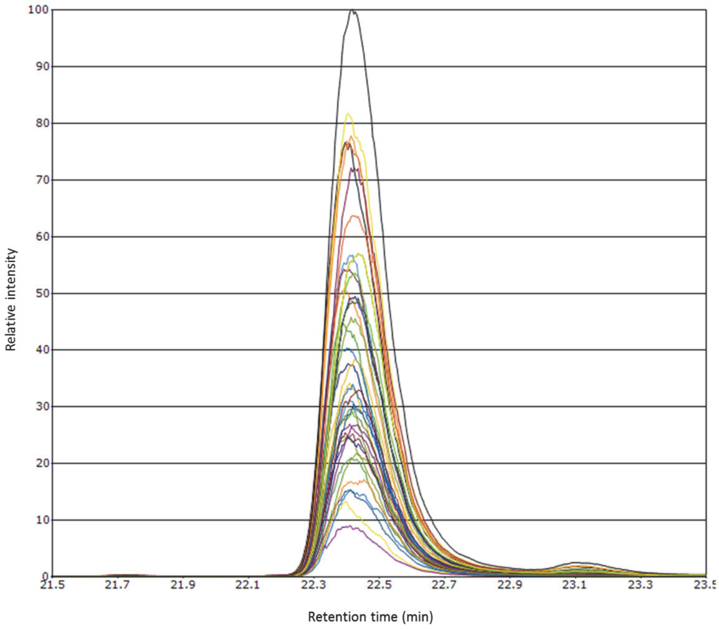


Figure 23: Chromatogram of 38 AALPAAFK isotopologues labeled with ^{13}C and ^{15}N isotopes.

Composition of the calibration mixture

In targeted experiments such as SRM or PRM, peptide quantification is based on the measurement of fragment ions. During the fragmentation process potential interferences can occur when precursor ions with similar m/z are co-isolated by the quadrupole and generate near-isobaric fragment ions [72]. Therefore, all reporter peptide isotopologues cannot be analyzed simultaneously because some peptides do not have sufficient molecular mass differences to avoid crosstalk.

It is necessary to determine which isotopologues can be mixed together in order to avoid crosstalk and interference. Figure 24 describes an example of three isotopologues AALPAAFK (m/z 404.761, z = 2+), AALPAAFK (m/z 405.263, z = 2+) and AALPAAFK (m/z 405.761, z = 2+) which cannot be analyzed simultaneously. Indeed, due to low mass difference (1 or 2 Da) between the three isotopologues their isotope distributions are overlapping and therefore the signal of the targeted peptide M is contaminated by the signals of the isotopes M-2, M-1, M+1 or M+2 of the other peptides.

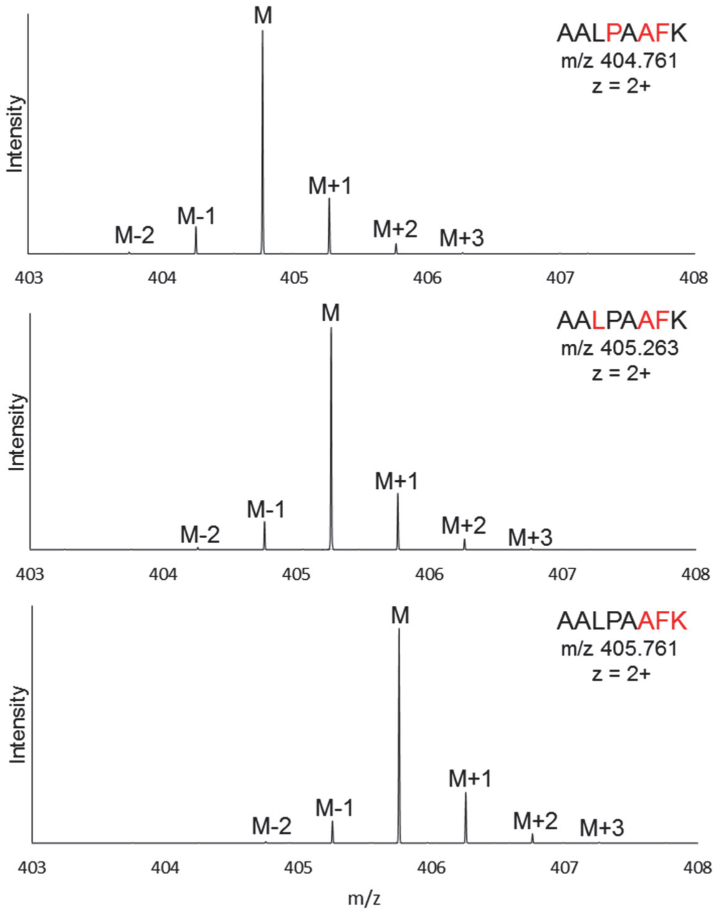


Figure 24: Isotopic distribution of three isotopologues of AALPAAFK.

For example during the simultaneous measurement of these three peptides the signal of AALPAAFK would be interfered by the signals of M+1 from AALPAAFK and M-1 from AALPAAFK. To ensure that there is no inferences between the constituents of the calibration mixture a minimum mass difference of 5 Da (2.5 Th for doubly charged precursor ions) between two adjacent isotopologues was fixed. This criteria allowed to define a set of ten isotopologues which can be analyzed simultaneously (Table 5).

Table 5: List of the ten AALPAAFK isotopologues composing the calibration mixture (labeled amino acids in red).

Sequence	Molecular mass	m/z (+2)	Δm
AALPAAFK	787.459	394.737	0
AALPAAFK	793.473	397.744	6
AALPAAFK	798.483	400.249	11
AALPAAFK	803.488	402.751	16
AALPAAFK	808.511	405.263	21
AALPAAFK	813.515	407.765	26
AALPAAFK	818.532	410.273	31
AALPAAFK	823.536	412.775	36
AALPAAFK	828.546	415.280	41
AALPAAFK	834.560	418.287	47

To confirm that the ten selected precursors for targeted experiments do not interfere, each isotopologue was analyzed individually on a Q-Exactive instrument. The reconstructed MS1 is presented in Figure 25. This analysis of the ten isotopologues validated the choice of a minimum of 5 Da mass difference between calibration constituents. The non-overlap of isotopic distributions allowed the simultaneous measurements of these ten isotopologues in a single LC-MS analysis. Concerning the isotopic distribution of the selected isotopologues we observed a profile change when the heavy stable isotope label increased. For the unlabeled peptide, the second most intense isotope is M+1. When the isotope labeling increases, the proportion of the isotope M+1 decreases gradually in benefit of the isotope M-1. For the isotopologue AALPAAFK (Δm=31) a profile inversion is observed, the isotope M-1 becomes the second most abundant isotope and its proportion continues to rise when the number of stable isotope labeled amino acids increases. This phenomenon is due to the isotope purity of the synthetic stable isotope labeled amino acids (99%). By increasing the number of labeled amino acids in a peptide the number of non-labeled atoms (1%) contributes to the increase of the isotopes M-1 and M-2. At the same time, when the number of labeled amino acids increases, the number of non-heavy isotope labeled amino acids decreases, contributing to the decrease of the isotope M+1 because of the contribution of the natural isotopes of the unlabeled amino acids (example for the natural isotope distribution of carbon: 98.9% ^{12}C and 1.1% ^{13}C).

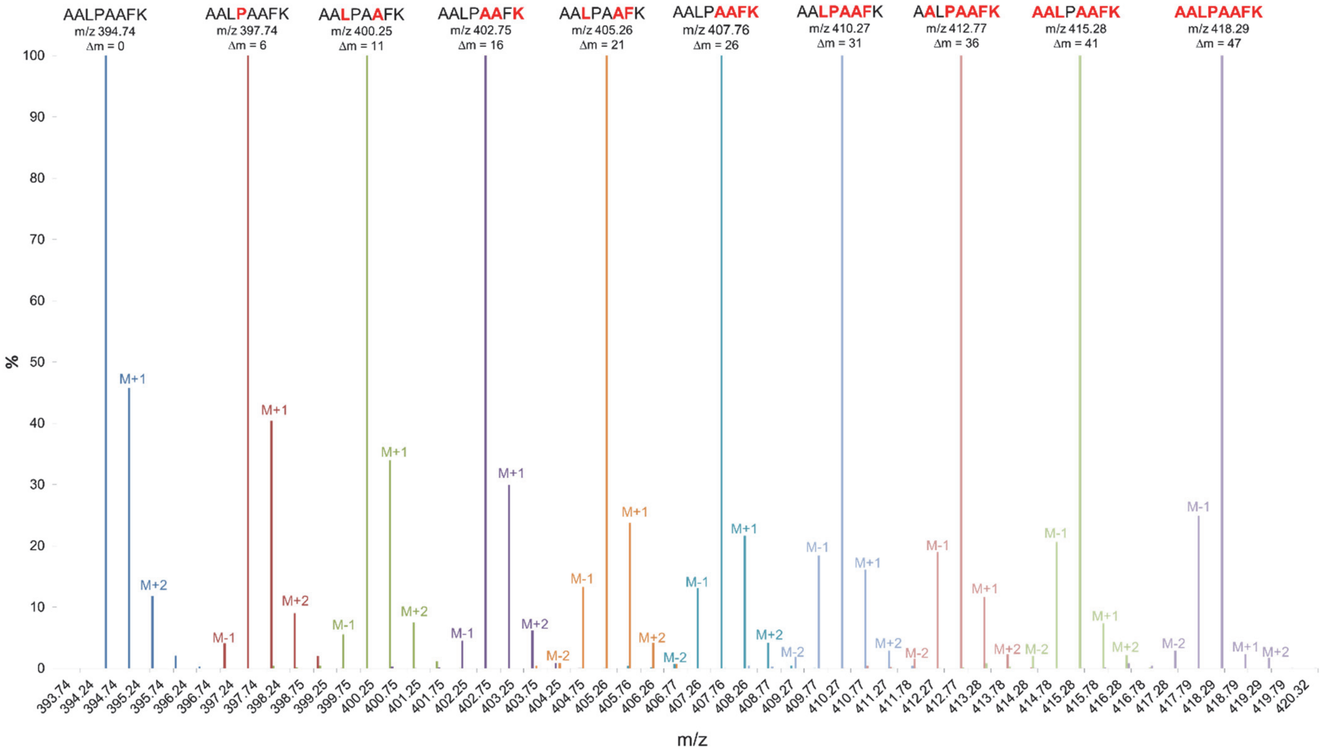


Figure 25: Reconstructed MS spectra from the individual analyses of the ten isotopologues constituting the calibration mixture (figure adapted from Duriez et *al.* Journal of Proteome Research, 2015; 2: 728-37).

Calibration curves in buffer, urine, and plasma samples

To be able to perform a reliable calibration of the cleavable reporter released during the trypsinization of the concatenated polypeptide, a linear response of the ten selected isotopologues, which constitute the calibration mixture, is required. To test this linear response, a mixture containing the ten istopologues, precisely quantified in various amounts (Table 6), was added in three different matrices: buffer, urine and plasma.

Table 6: Composition of the calibration mixture containing ten isotopologue reporter peptides in various amounts (table adapted from Duriez et *al.* Journal of Proteome Research, 2015; 2: 728-37).

Sequence	**Molecular mass**	**m/z (+2)**	**∆m**	**Concentration (amol/µL)**
AALPAAFK	787.459	394.737	0	30000
AALPAAFK	793.473	397.744	6	10000
AALPAAFK	798.483	400.249	11	3333
AALPAAFK	803.488	402.751	16	1111
AALPAAFK	808.511	405.263	21	370
AALPAAFK	813.515	407.765	26	123
AALPAAFK	818.532	410.273	31	41
AALPAAFK	823.536	412.775	36	14
AALPAAFK	828.546	415.280	41	5
AALPAAFK	834.560	418.287	47	2

Each isotopologue dilution curve in the three matrices was analyzed by SRM on a triple quadrupole mass spectrometer by monitoring transitions from y_2 to y_7 for the 10 isotopologues simultaneously. The dilution curves made from ten reporter isotopologues in various amounts prove the linearity of the measurements in the three different backgrounds (Figure 26). For each condition, LOQs were determined for the reporter peptides according to several criteria. First, the maximum CV allowed for the lowest concentration was 20% and second, the accuracy had to be between 80 and 120 %. In buffer, urine and plasma the LOQs were 5 amol, 40 amol and 125 amol (injected on column), respectively. The linearity of the isotopologue calibration curves in the different matrices indicate that the reporter peptide is not prone to ion suppression and that the transitions selected for SRM are not interfered by the different backgrounds. The reporter peptide calibration can thus be performed directly in the biological matrix.

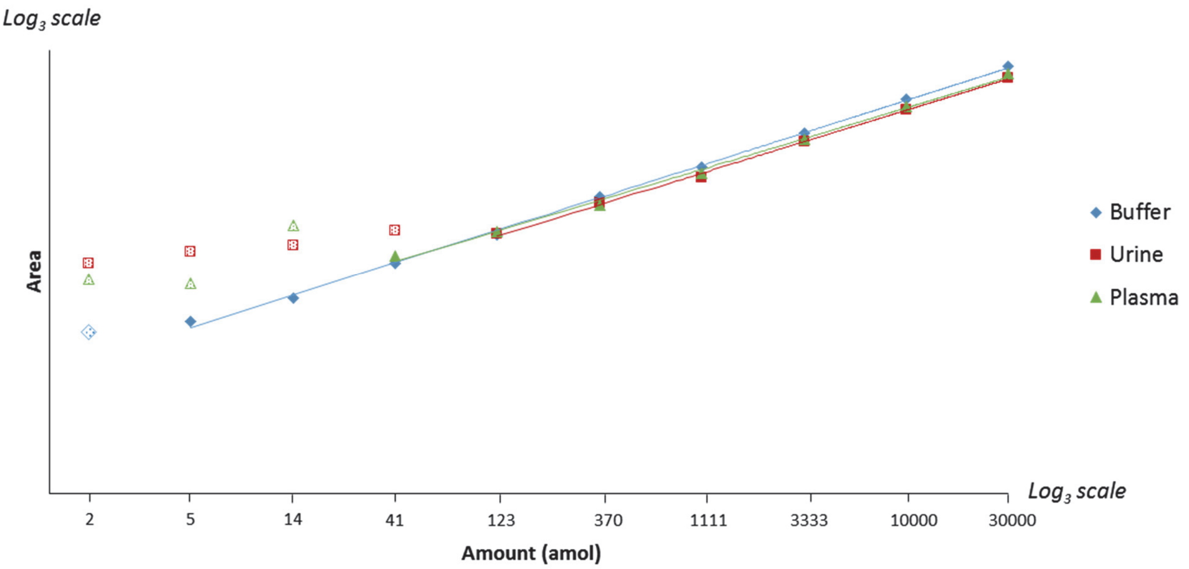


Figure 26: Calibration curves of the ten isotopologues analyzed in three different backgrounds: buffer, urine and plasma (figure adapted from Duriez et *al.* Journal of Proteome Research, 2015; 2: 728-37).

Multiplexing degree of the concatenated polypeptides

The degree of multiplexing of the method corresponds to the maximum number of isotopically-labeled signature peptides which can be calibrated simultaneously. A minimum of four isotopologues is required to compose the calibration curve. Thus, with ten isotopologues available the multiplexing degree is six, meaning that six signature peptides fused with six different cleavable reporter peptides can be calibrated concomitantly.

5. Quantification methods using concatenated polypeptides

Peptide quantification based on concatenated polypeptides used as internal standards can be performed in two different ways. First, by adding the calibrated isotopically-labeled peptide internal standards to the biological sample digest obtained after an external quantification. Second, by generating the signature peptide standards *in-situ* by adding the polypeptides directly to the biological sample before proteolysis. In this case the calibration of the isotopically-labeled peptide internal standards will be carried out directly in the biological sample digest by measuring all the reporter peptide intensities during the LC-MS analysis.

Peptide quantification using external calibration

In the quantitative strategy involving an external calibration of the standard (Figure 27), the concatenated polypeptide is first digested in buffer with trypsin releasing the isotopically-labeled signature peptide and the cleavable reporter in equimolar amounts. Second, the signature peptide standard is calibrated by LC-MS analysis via the accurate determination of the reporter peptide amount using a calibration mixture of isotopologue reporter peptides. Finally, the endogenous peptide quantification is performed by targeted LC-MS analyses using a known amount of the calibrated isotopically-labeled signature peptide internal standard which was added to the peptide mixture as with the conventional calibrated isotopically-labeled peptides. The individual calibrated isotopically-labeled signature peptides can be stored after use and reused later after a prior recalibration procedure.

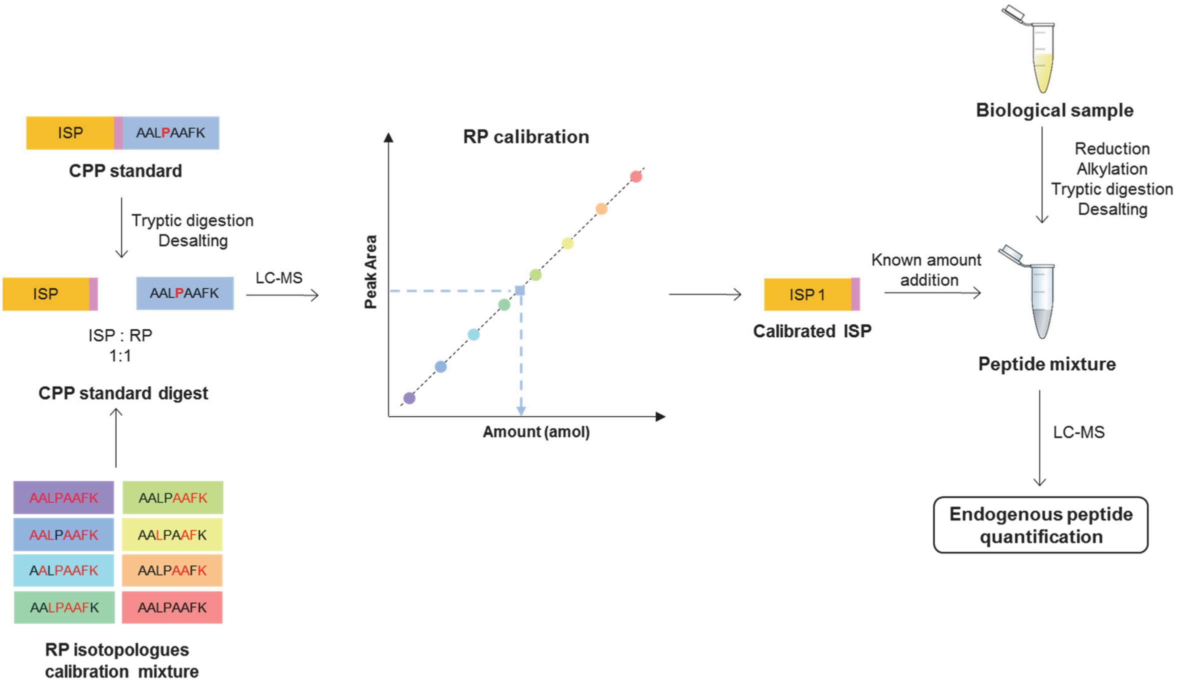


Figure 27: Protein quantification strategy using the concatenated polypeptides involving an external calibration of the released isotopically-labeled signature peptide.

Peptide quantification using internal calibration

In the quantitative strategy involving an internal calibration of the standard (Figure 28), the concatenated polypeptide follows the same process as the biological sample. First, the polypeptide is added to the biological sample. Second, the sample containing the standard is enzymatically digested with trypsin to generate peptides. Third, the digest is supplemented with the isotopologue reporters for the calibration. The amount of released isotopically-labeled signature peptide is determined using the released reporter peptide. The endogenous peptide is simultaneously quantified during the same LC-MS run. The internal calibration of the isotopically-labeled signature peptide standards is more straightforward than the external calibration but only six endogenous peptides can be quantified per biological sample as four isotopologue reporters are needed for the calibration curve.

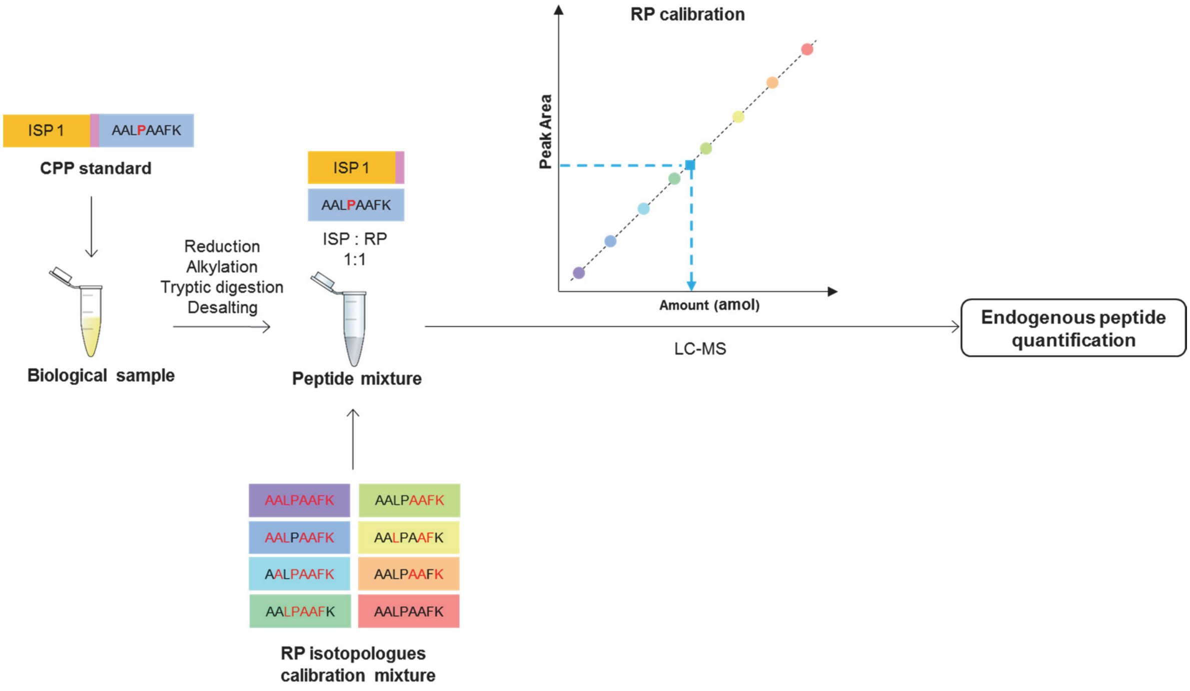


Figure 28: Protein quantification strategy using the concatenated polypeptides involving an internal calibration of the released isotopically-labeled signature peptide.

Comparison of external and internal calibration applied to serum amyloid A in plasma

Serum amyloid A (SAA) is a plasma protein belonging to the family of apolipoproteins secreted by the liver in response to inflammation induced by infection or tissue injury [114]. A high level of SAA in plasma can reflect a chronic inflammatory state as it can be the case during the development of a lung cancer [115-118]. An evaluation of the concordance of the quantitative results obtained with signature peptides calibrated internally and externally was conducted with the precise quantification of SAA protein in a lung cancer plasma sample. This comparison was carried out using two concatenated polypeptide standard isotopologues, GPGGVWAAEAISDARAALPAAFK and GPGGVWAAEAISDARAALPAAFK. The first one was used to perform the quantification of SAA using the internal calibration approach and the second one using the external calibration approach (Figure 29).

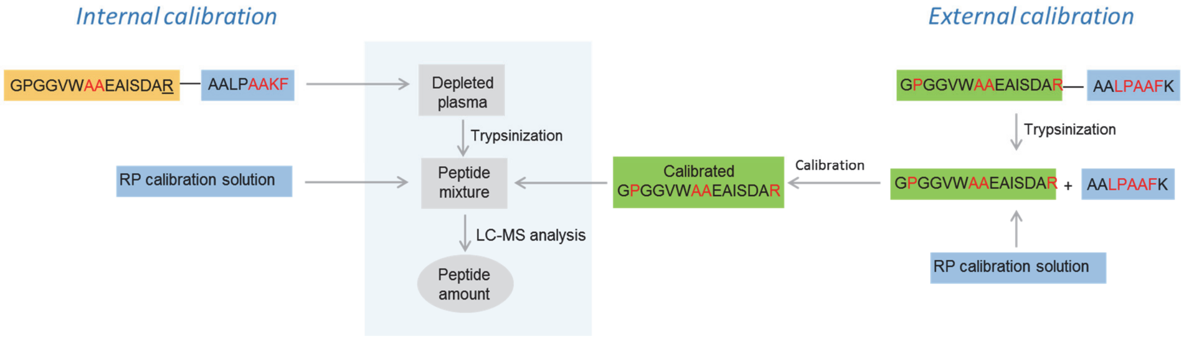


Figure 29: Workflows for serum amyloid A quantification in plasma coming from a patient diagnosed with lung cancer using the cleavable reporter peptide strategy involving an internal and an external calibration of the released isotopically-labeled signature peptide (figure adapted from Duriez et *al.* Journal of Proteome Research, 2015; 2: 728-37).

In this experiment, GPGGVWAAEAISDARAALPAAFK was added to the depleted plasma sample and was enzymatically digested in the plasma matrix. In parallel, GPGGVWAAEAISDARAALPAAFK was digested alone in buffer and the released ISP was calibrated. After that, a known amount of GPGGVWAAEAISDAR (externally calibrated) and the reporter calibration mixture were supplemented into the mixture of plasma peptides. Finally, in a single LC-MS run GPGGVWAAEAISDAR and GPGGVWAAEAISDAR were respectively calibrated and recalibrated (to evaluate the effect of the matrix on the calibration process). Figure 30 presents the quantitative values for the endogenous peptide GPGGVWAAEAISDAR determined using the CPP digested and calibrated in plasma (1), the CPP digested and calibrated in buffer (2) and the CPP digested in buffer and recalibrated in plasma (3). For the three calibration conditions of the internal standard, very similar concentrations were measured for the endogenous SAA peptide demonstrating that peptide quantification using isotopically-labeled peptide calibrated internally or externally can be performed in an equivalent way and also that the presence of a complex matrix does not bias the SIL peptide calibration at least for relatively abundant proteins.

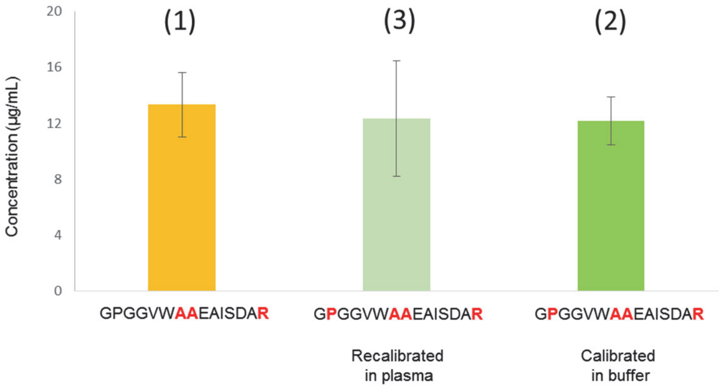
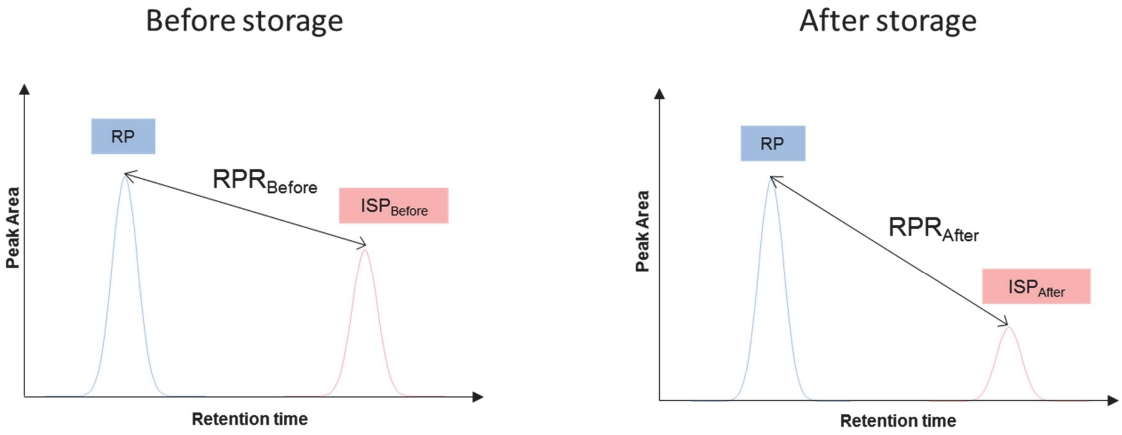


Figure 30: Quantitative values of serum amyloid A measured in plasma using the surrogate peptide GPGGVWAAEAISDAR for the different concatenated polypeptide strategies. (1) GPGGVWAAEAISDAR internally calibrated in plasma, (2) GPGGVWAAEAISDAR externally calibrated in buffer and (3) GPGGVWAAEAISDAR externally calibrated in buffer and recalibrated in plasma (figure adapted from Duriez et *al.* Journal of Proteome Research, 2015; 2: 728-37).

6. Recalibration procedure of the ISP internal standard

During the tryptic digestion of a concatenated polypeptide, the signature peptide and the cleavable reporter are released in equimolar amounts. Each polypeptide digest is characterized by the released peptide ratio (RPR) which is the ratio of intensities between the isotopically-labeled signature peptide and the cleavable reporter. The recalibration procedure of signature peptide internal standards of one concatenated polypeptide digest is based on the comparison of the two RPRs before and after storage allowing to estimate the loss of isotopically-labeled signature peptide (the reporter amount is considered as invariant) (Figure 31).



$$\text{Amount}_{\text{ISP After}} = \frac{\text{RPR}_{\text{Before}}}{\text{RPR}_{\text{After}}} \times \text{Amount}_{\text{ISP Before}}$$

Figure 31: Recalibration procedure of isotopically-labeled signature peptide internal standards based on the comparison of the RPRs before and after storage.

The isotopically-labeled signature peptide recalibration approach is directly linked to the stability of the reporter peptide. To illustrate this point, analyses were performed to measure the stability of the reporter peptide stored in water in a low binding plastic tube container after four freeze-thaw cycles. A high recovery rate of about 99% for the reporter was observed after the four freeze-thaw cycles allowing to consider the reporter as reference for the recalibration of the isotopically-labeled signature peptide after storage.

To highlight the necessity of systematic standard peptide recalibration, an evaluation of the stability of concatenated polypeptide digests after four storage freeze-thaw cycles was performed. Figure 32 presents the signature peptide recovery after four freeze-thaw cycles for five polypeptide digests. For these examples, partial losses of the isotopically-labeled signature peptides were observed ranging from 19 to 87%, probably due to non-specific hydrophobic interactions between the peptides and the container surfaces. The most important loss was observed after the first cycle. These results show that storage affects the isotopically-labeled signature peptide amount and consequently can induce biases during protein quantifications by overestimating the amount of endogenous peptides.

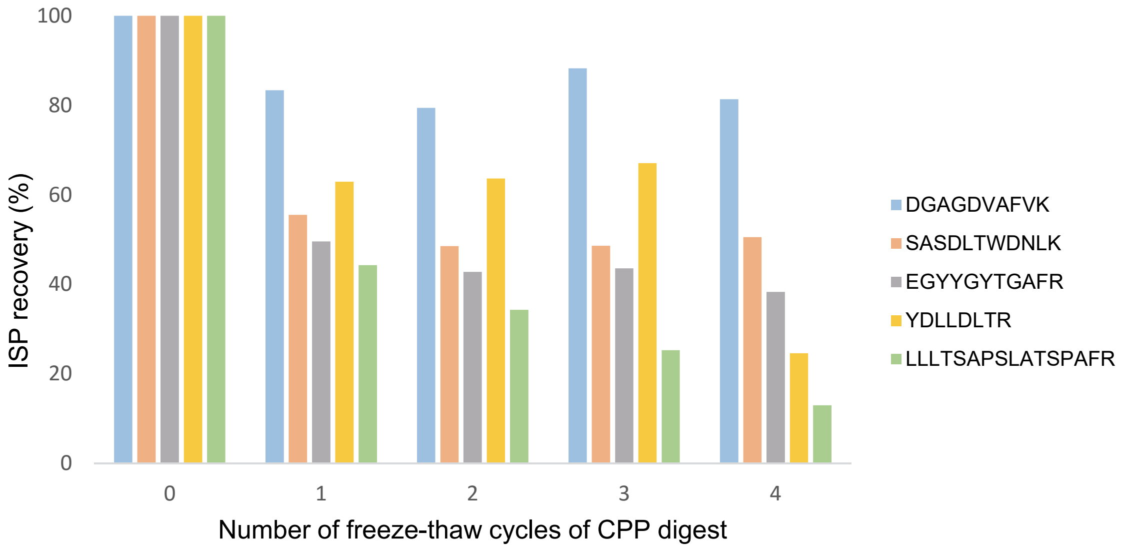


Figure 32: Isotopically-labeled signature peptide recovery after freeze-thaw cycles for five cleavable polypeptide digests (figure adapted from Duriez et *al.* Journal of Proteome Research, 2015; 2: 728-37).

The loss of isotopically-labeled signature peptide standards after freeze-thaw cycles is usually not taken into consideration in most studies but the present data clearly indicates the need of recalibration of the internal standards before each use to achieve accurate protein quantification. The approach based on concatenated polypeptides as internal standards allows systematic recalibration of signature peptide standards before spiking into biological samples.

7. Concatenated polypeptides with two signature peptides

The use of cleavable polypeptides as internal standards for peptide quantification allows for generating more precise and accurate results, but its multiplexing capacities are limited due to the number of isotopologues available. Indeed, ten isotopologue reporters can be analyzed together, four of them are dedicated to the calibration mixture so a maximum of six standards can be calibrated at the same time. In order to extend the multiplexing capacities of the approach and also to reduce the number of polypeptide standards to handle, especially for the quantification of panels of proteins, a second generation of concatenated polypeptide standards containing a cleavable reporter peptide and several isotopically-labeled signature peptides of interest was explored. Quantification using polypeptides containing multiple signature peptides is highly dependent on the proteolysis efficiency of the different cleavage sites. Indeed, during the proteolysis of the standard, all the peptide constituents are expected to be released in an equimolar amount allowing the simultaneous calibration of all the polypeptide constituents using a single reporter (after the prior addition of the isotopologue calibration mixture). To evaluate the digestion efficiency of this second generation of concatenated polypeptide standards, twenty three polypeptides containing two signature peptides and the cleavable reporter AALPAAFK were used, as described Figure 33.

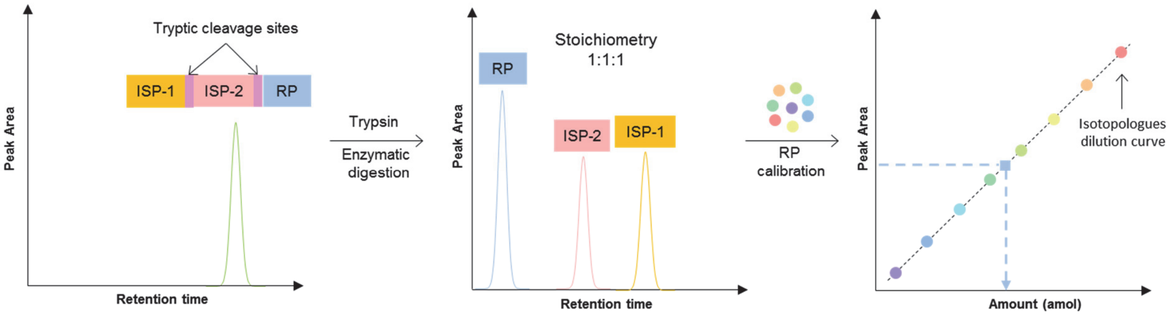


Figure 33: Description of the strategy of the concatenated polypeptide containing two ISPs concatenated with a single cleavable reporter.

In this case, the concatenated polypeptide contains two tryptic cleavage sites which separate its three constituents: the ISP-1, the ISP-2 and the reporter. Each polypeptide was digested individually with trypsin, separated by liquid chromatography and analyzed by mass spectrometry in full scan mode. The presence of missed cleavages in the digests was investigated. The equimolarity of the three expected released products is essential to the quantification process. Results of the identified species from the 23 polypeptides are summarized in Table 7. In this table, the tick mark indicates that the corresponding peptide has been observed in the corresponding polypeptide digest. In all digests the three constituents were identified, but in a vast majority, large peptides containing one or/and two missed cleavages were observed. These results indicate that the use of concatenated polypeptides containing several signature peptides leads the generation of ISP-1, ISP-2 and reporter in different amounts which affects the accuracy of the quantification experiment.

Table 7: Identified missed cleaved species in twenty three concatenated polypeptide digests containing two isotopically-labeled signature peptides.

Concatenated polypeptide	1 missed cleavage		2 missed cleavages
ISP-1 -- ISP-2 -- RP	**ISP-1 -- ISP-2**	**ISP-2 -- RP**	**ISP-1 -- ISP-2 -- RP**
SASDLTWDNLKDGAGDVAFVKAALPAAFK	✓	✓	
EGYYGYTGAFRDGAGDVAFVKAALPAAFK	✓	✓	
TAGIQIVADDLTVTNPKLGANAILGVSLAASRAALPAAFK			✓
GVIFYESHGKSIGGEVFIDFTKAALPAAFK		✓	✓
VVGLSTLPEIYEKSIGGEVFIDFTKAALPAAFK		✓	
DSTIQVVENGESSQGRDWVSVVTPARAALPAAFK	✓		
SGSVIDQSRDWVSVVTPARAALPAAFK		✓	✓
SDLAVPSELALLKELSEALGQIFDSQRAALPAAFK	✓	✓	
LADGGATNQGRELSEALGQIFDSQRAALPAAFK	✓	✓	✓
YGFIEGHVVIPRFAGVFHVEKAALPAAFK		✓	
ALSIGFETCRFAGVFHVEKAALPAAFK	✓	✓	
ALSIGFETCRYGFIEGHVVIPRAALPAAFK	✓	✓	
VSTLPAITLKLVDQNIFSFYLSRAALPAAFK		✓	✓
YSQAVPAVTEGPIPEVLKLVDQNIFSFYLSRAALPAAFK	✓		✓
YDLLDLTRLLLTSAPSLATSPAFRAALPAAFK			
SDVFEAWRLLLTSAPSLATSPAFRAALPAAFK			
NTEISFILGQEFDEVTADDREVGVGFATRAALPAAFK	✓		
STITLDGGVLVHVQKEVGVGFATRAALPAAFK	✓	✓	✓
YDEEFASQKAWTDVLPWKAALPAAFK			✓
LTIESTPFNVAEGKSDLVNEEATGQFRAALPAAFK	✓	✓	
CETQNPVSARLTIESTPFNVAEGKAALPAAFK	✓	✓	
NLLSVAYKAVTELNEPLSNEDRAALPAAFK		✓	✓
VISSIEQKAVTELNEPLSNEDRAALPAAFK		✓	

8. Conclusion

In the context of clinical studies such as for the determination of new cancer biomarker targets, precise, accurate and reliable quantitative methods are required. This remains very challenging. In bottom-up approaches, accurate protein quantification is commonly performed using high purity calibrated stable isotope-labeled peptides. With this type of standard, the quantification process is only based on the nominal amount provided by the supplier and does not take into account unexpected events such as an incomplete solubilization, non-specific adsorption on hydrophobic surfaces or aggregation which can induce a quantification bias. The developed approach using concatenated polypeptides is able to address the limitations of common calibrated stable isotope-labeled peptides due to its capability to calibrate the internal standards before the quantitative analyses.

Chapter II: Development of non-tryptic digestion methods

This chapter describes experiments reported in the article *Evaluation of alternative enzymes to trypsin for enhancing proteomics analysis* by S. Trévisiol, D. Ayoub, A. Lesur, S. Gallien and B. Domon, recently submitted for publication (Annex 4).

1. Introduction

The large majority of bottom-up experiments uses trypsin to perform the proteolysis. Protein digestions with trypsin result in a high sequence coverage and consequently in a large number of protein identifications after LC-MS analyses of complex biological samples. However, an uneven distribution of lysine and arginine residues in proteins [119] causes some parts of protein sequences to be unaccessible to trypsin digestion. This may be problematic because those missing parts may contain crucial information for disease understanding. For instance in the context of protein isoforms quantification, trypsin does not always produce the correct peptide sequences to distinguish the different protein forms [118]. Moreover, tryptic digestion generates large numbers of short peptides resulting in an increased background complexity which may induce signal interferences during the MS acquisition process [120]. As alternative to trypsin different proteases, such as Arg-C, Lys-N, Lys-C, Glu-C or Asp-N, are commonly used especially to improve protein sequence coverage [44, 121]. In this chapter the use of alternative enzymes was investigated in order to study their effects on the resulting background complexity and protein coverage. Furthermore, a systematic evaluation of the collision energy effect on the fragmentation pattern was carried out for different types of peptides.

2. Is there a need for alternative proteolytic enzymes?

Trypsin has been extensively used in many proteomic experiments but the consequences of its specificity *i.e.,* the generation of samples with increased complexity and a reduction of protein sequence coverage makes it unsuitable for certain types of proteomics experiments. The use of alternative enzymes having different cleavage sites may overcome these limitations.

Reduction of peptide density for LC-MS analyses

Biological samples as for example used in clinical studies are already a very complex mixture of proteins which after proteolytic digestion into peptides will dramatically increase sample complexity. Depending on the enzyme used different amounts of peptides, resulting in different backgrounds, are generated due to their cleavage specificities. Background complexity (number of entities) has a direct impact on the analytical performance of the mass spectrometer. The more complex the background is, the more potential interferences can occur during peptide detection and quantification.

In-silico digestion of the human proteome

To estimate the degree of complexity of the backgrounds generated by different enzymes an *in-silico* digestion of the human proteome (NeXtProt database, version 2014-05-27) was performed with five commonly used proteases *i.e.,* trypsin, Lys-C, Arg-C, Lys-N, Asp-N and Glu-C. The theoretical number of non-redundant peptides and the average molecular masses for peptides containing at least five amino acid residues and a molecular mass below 5 kDa are presented in figures 34A and 34B, respectively. Figure 34 shows that the enzymes Lys-C, Arg-C, Lys-N and Glu-C (E) which have a single specific cleavage site generate fewer peptides compared to enzymes that have two cleavage sites such as trypsin, Asp-N and Glu-C (E/D). Trypsin produces roughly twice the amount of peptides than Lys-C and Arg-C with the peptides generated by the latter two having on average a 1.5 fold increased mass. Generating a smaller number of larger peptides is beneficial to LC-MS analysis as the peptide density is lower which reduces the gap between the number of components and the peak capacity of the analytical system.

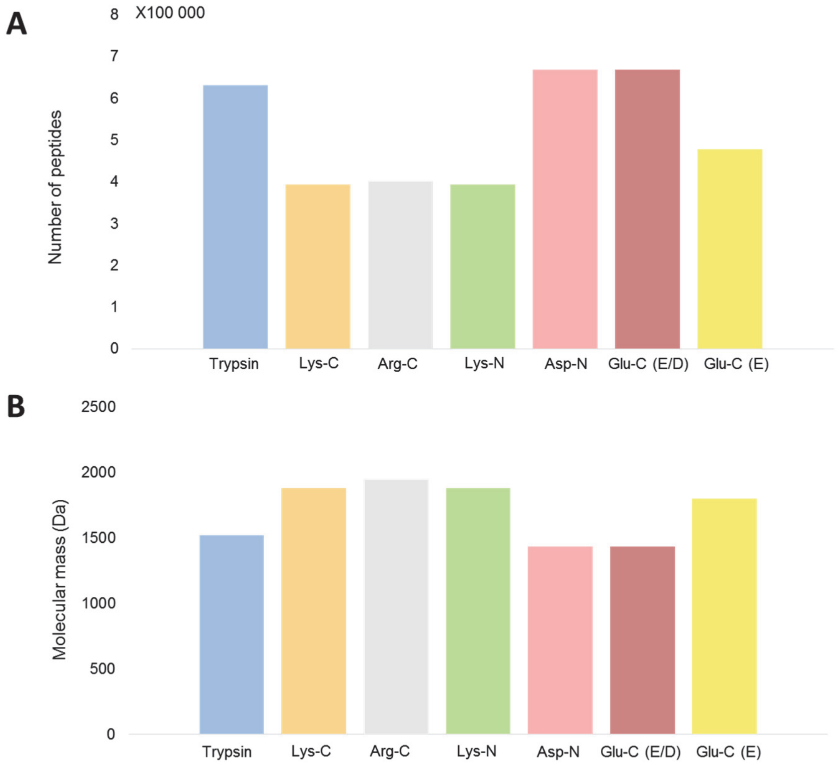


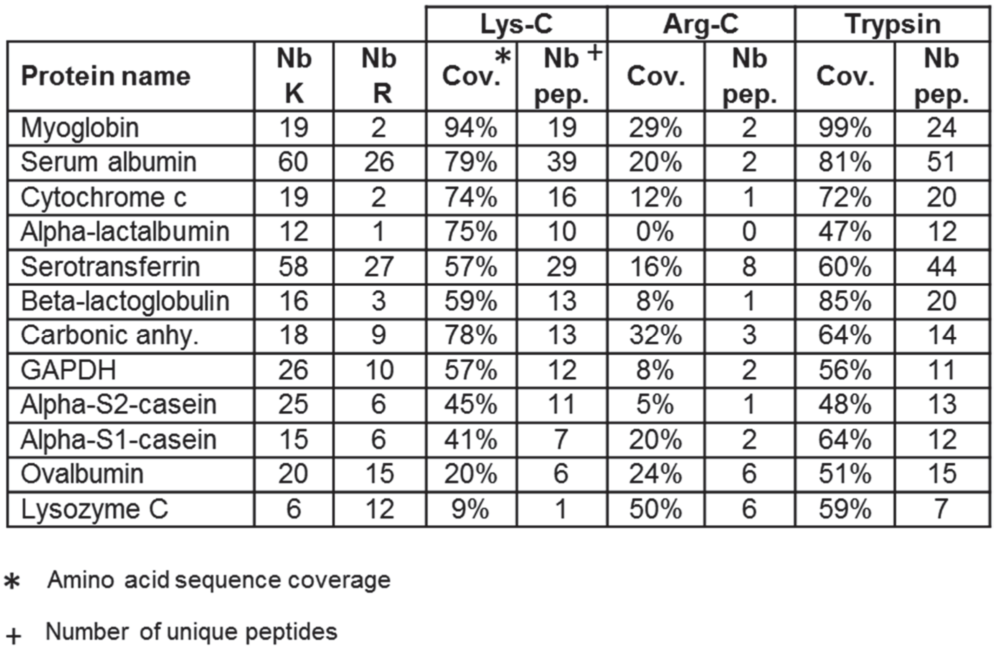
Figure 34: Theoretical number (A) and average molecular mass (B) of non-redundant peptides with a mass < 5000 Da but containing at least five amino acid residues (*in-silico* digestion of the human proteome with five different proteases).

Approaches to generate large peptides

A general strategy to decrease the complexity of the background is to cleave proteins into larger peptides than those obtained with trypsin. Different approaches were evaluated, the first involving the use of alternative enzymes and the second, the chemical modification of lysine residues (making these resistant to trypsin digestion) followed by a tryptic digestion.

Lys-C and Arg-C, unlike trypsin, are proteases which have a single specific cleavage site after lysine or arginine residues, respectively. After enzymatic digestion of a complex sample with either Lys-C or Arg-C consequently fewer, but larger, peptides are generated which may result in a significant reduction in sample complexity. This reduction in sample complexity may improve the identification and quantification of potential biomarkers at very low concentration in complex matrices such as plasma. To evaluate the ability of Lys-C or Arg-C to generate large peptides, a mixture of twelve proteins was digested in parallel with the two enzymes, and compared to a tryptic digestion of the same sample. The mass spectrometric analysis of each protein digest was performed, after a liquid chromatography separation using a C_{18} analytical column, in DDA mode on a Q-Exactive mass spectrometer. Peptide identifications were performed by querying Mascot using a restricted database containing the sequences of the twelve proteins and by a manual search. For the three digestions, the number of identified peptides and the protein sequence coverage are presented in Table 8.

Table 8: Protein coverage obtained for twelve proteins in three different digests i.e. Lys-C, tryptic and Arg-C.



			Lys-C		Arg-C		Trypsin	
Protein name	**Nb K**	**Nb R**	**Cov.** *	**Nb** + **pep.**	**Cov.**	**Nb pep.**	**Cov.**	**Nb pep.**
Myoglobin	19	2	94%	19	29%	2	99%	24
Serum albumin	60	26	79%	39	20%	2	81%	51
Cytochrome c	19	2	74%	16	12%	1	72%	20
Alpha-lactalbumin	12	1	75%	10	0%	0	47%	12
Serotransferrin	58	27	57%	29	16%	8	60%	44
Beta-lactoglobulin	16	3	59%	13	8%	1	85%	20
Carbonic anhy.	18	9	78%	13	32%	3	64%	14
GAPDH	26	10	57%	12	8%	2	56%	11
Alpha-S2-casein	25	6	45%	11	5%	1	48%	13
Alpha-S1-casein	15	6	41%	7	20%	2	64%	12
Ovalbumin	20	15	20%	6	24%	6	51%	15
Lysozyme C	6	12	9%	1	50%	6	59%	7

* Amino acid sequence coverage

+ Number of unique peptides

The comparison of the tryptic and the Lys-C digestions confirmed the generation of a smaller number of peptides with Lys-C. Indeed, for the majority of proteins, similar protein coverages were observed for both trypsin and Lys-C digestions albeit with lower number of peptides generated during Lys-C proteolysis. This reflects the fact that Lys-C produces larger peptides as compared to trypsin. For some proteins, as for example Lysozyme C and Ovalbumin, lower protein coverages were noticed for Lys-C in comparison to the tryptic digestion. This difference was not due to a poor digestion efficiency of Lys-C but to the amino acid composition of these proteins i.e., the number of lysine residues and/or their distribution. Regarding the Arg-C proteolysis, the protein mixture used in this experiment was not really suitable to evaluate if the enzyme generates large peptides due to the low frequency of arginine residues in the different proteins. Only Lysozyme C which contains more arginine than lysine residues obtained a protein coverage similar to trypsin.

Due to the large scale utilization of trypsin the cost of this enzyme is lower than the cost of other enzymes. In order to generate larger peptides without the use of the expensive alternative enzymes, a chemical modification of lysine residues in proteins followed by tryptic digestion to generate Arg-C-like peptides was evaluated. The chemical modification of the lysine residues makes them unaccessible to trypsin (by neutralizing the positive charge on the amine group) (Figure 35).

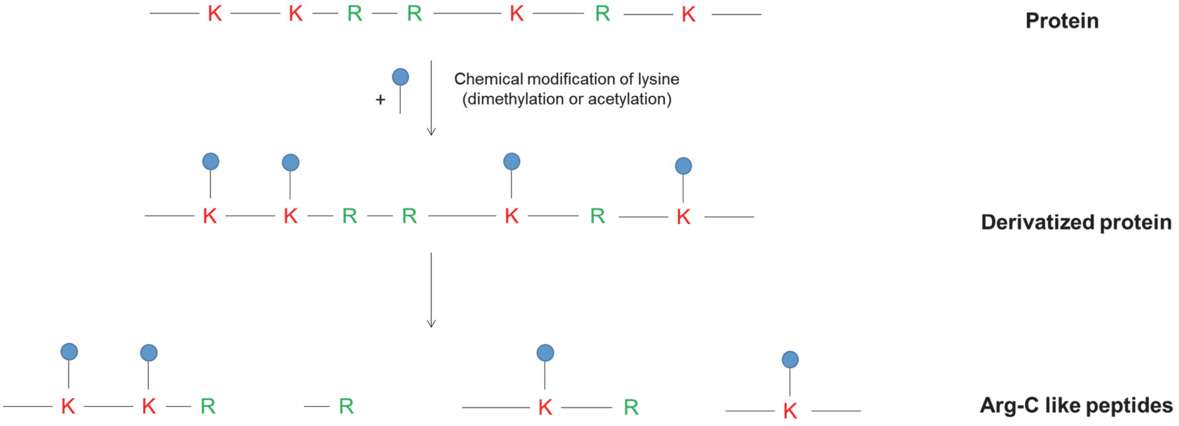

K
K
R
R
K
R
K
Protein
+
Chemical modification of lysine
(dimethylation or acetylation)
Derivatized protein
Arg-C like peptides

Figure 35: Workflow to generate Arg-C like peptides.

Two different chemical reactions were considered, the dimethylation and the acetylation using formaldehyde (CH_2O) and sulfo-N-hydroxysuccinimide acetate (sulfo-NHS acetate), respectively. To evaluate the efficiency of the chemical modification, the reactions were performed on melittin, a 26 amino acid polypeptide (GIGAVLKVLTTGLPALISWIKRKRQQ-NH2) which is a component of bee venom.

Dimethylation of lysine residues is a two-step reductive alkylation reaction of free amine groups with the aldehyde group to form a Schiff base, which can be reduced to an amine group [122-125]. Formaldehyde is used as the alkylating agent and reduction is performed with sodium cyanoborohydride ($NaBH_3CN$) (Figure 36).

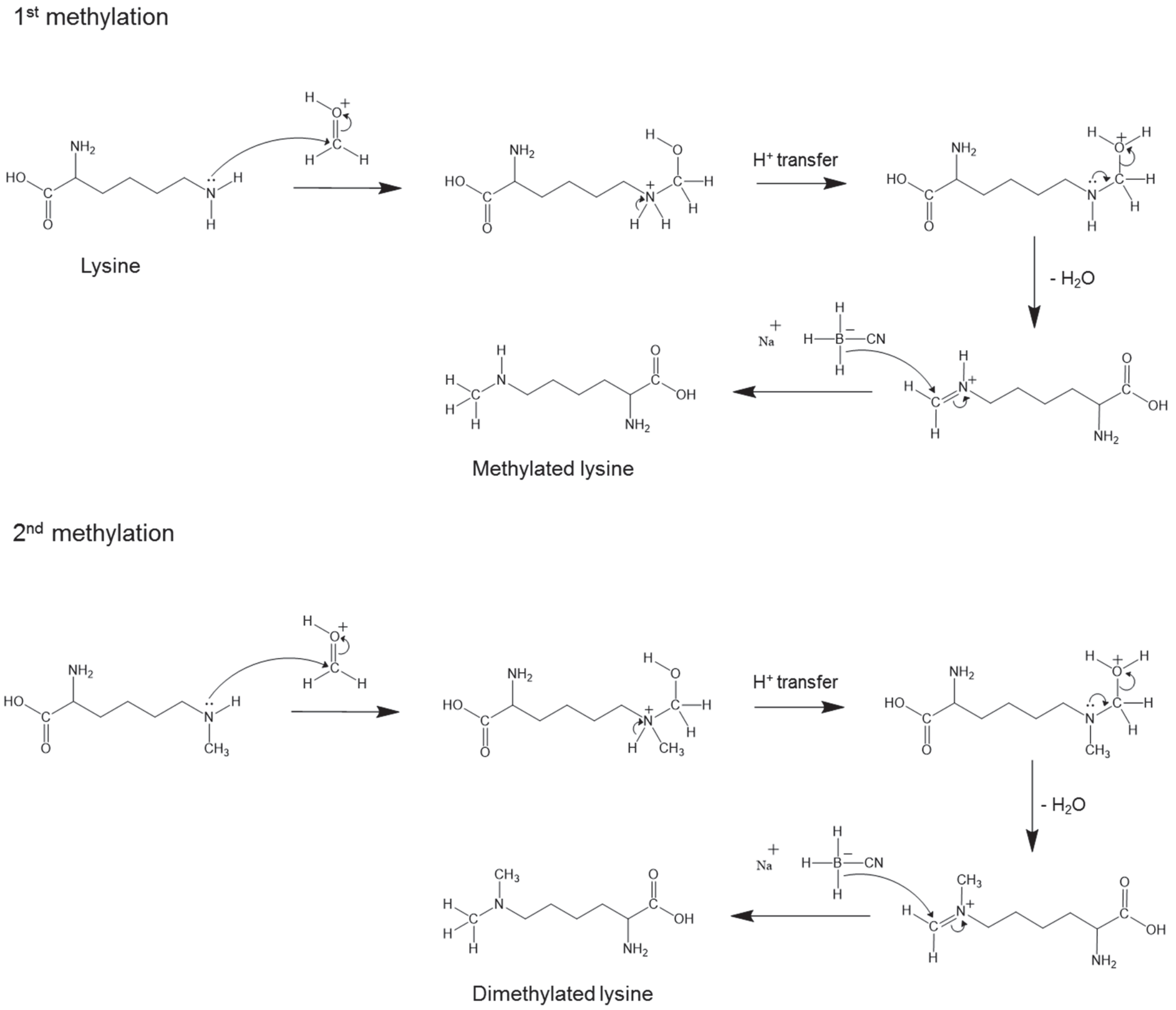


Figure 36: Mechanism of dimethylation of lysine residues using CH_2O as alkylating agent and sodium $NaBH_3CN$ as reducing agent.

Melittin contains four amine groups, one at the N-terminal position and one per lysine residue. So four sites were expected for dimethylation. After the dimethylation modification the polypeptide was analyzed in MS1 by LC-MS. The MS1 spectrum of the five times charged dimethylated protein (Figure 37) showed that the chemical reaction was not quantitative. Indeed, the reaction should have resulted in eight methylation modifications (+112 Da) of the protein, but, although this molecule was predominantly produced substantial amounts of side products, corresponding to incomplete methylation, were also observed.

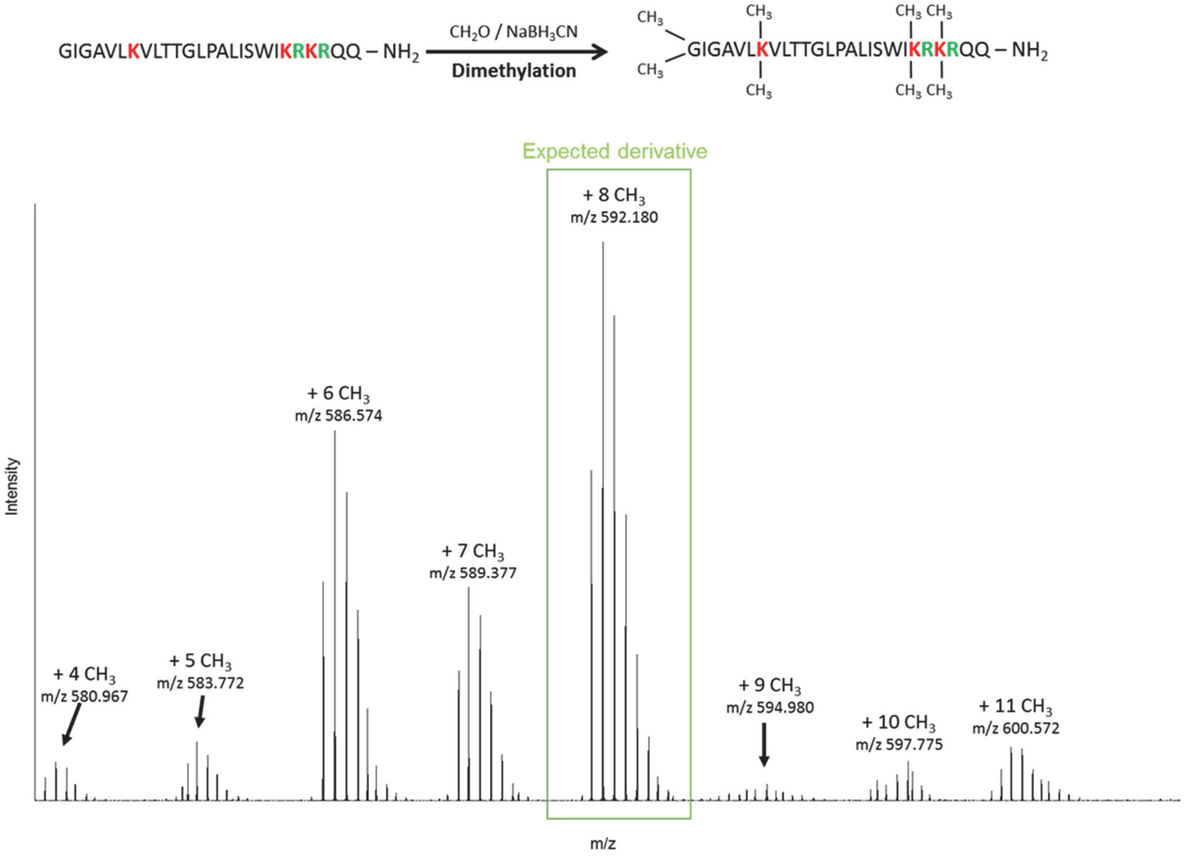


Figure 37: MS1 spectrum of the five times charged dimethylated melittin.

The chemical acetylation of lysine residues introduces an acetyl group on the primary amino groups of lysine residues and the N-terminus as a result of the nucleophilic attack of the amine nitrogen to the electrophilic carbonyl group [126-128] of sulfo-N-hydroxysuccinimide acetate ester (Figure 38).

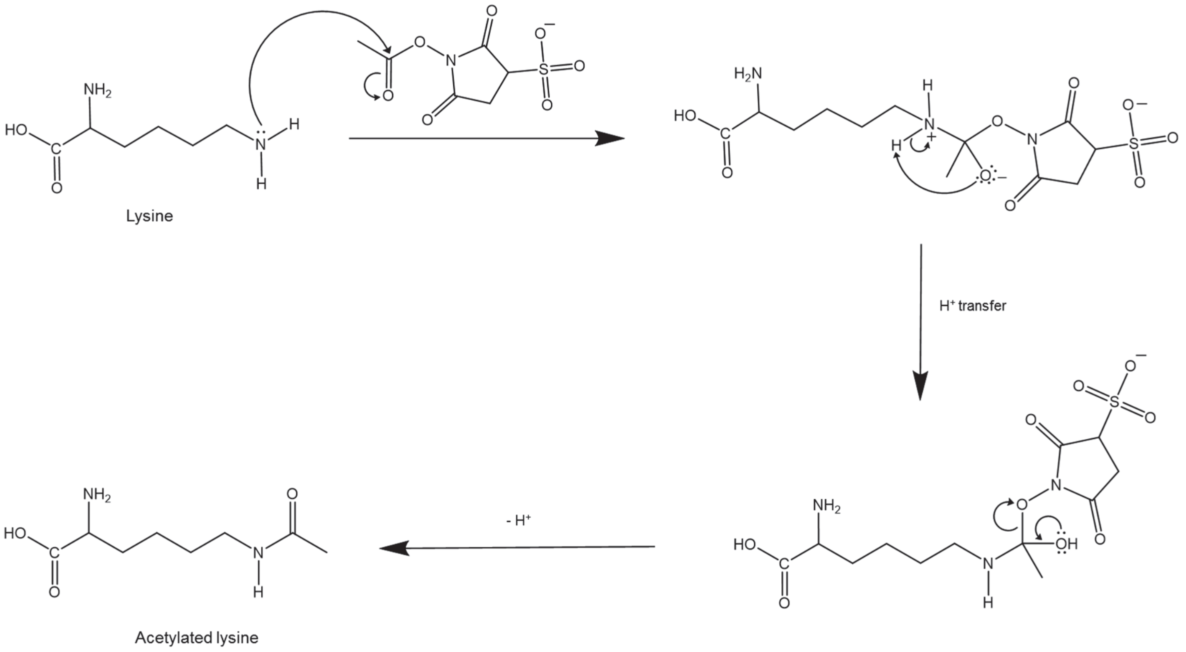


Figure 38: Mechanism of acetylation of a lysine residue using sulfo-NHS acetate as electrophilic agent.

The evaluation of the acetylation efficiency was performed also on melittin with four sites available for chemical modification. After acetylation the protein was analyzed in MS1 by LC-MS. Figure 39 presents the extracted ion chromatogram traces of the triply charged di-, tri-, and tetra-acetylated melittin illustrating the same behavior as observed for dimethylation *i.e.,* a non-quantitative chemical reaction. Indeed for melittin the tetra-acetylated (+168 Da) form was expected but the di- and tri- acetylated forms were also observed despite the excess of reagent employed.

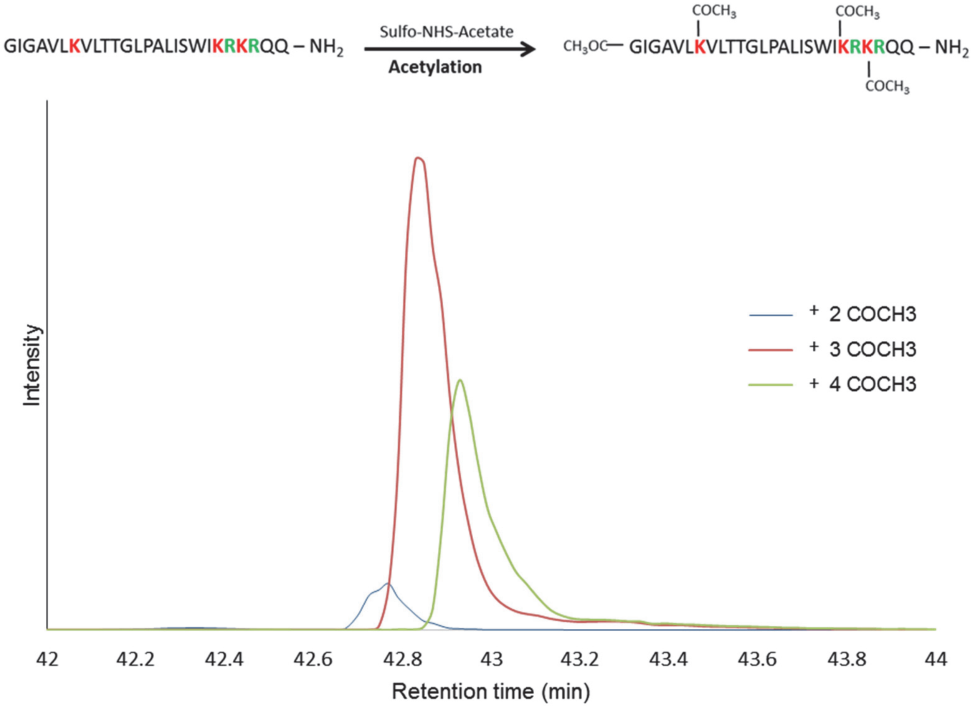


Figure 39: LC-MS extracted ion chromatogram traces of the triply charged di- (blue), tri- (red), and tetra-acetylated (green) melittin.

The main objective of the generation of Arg-C-like peptides using a chemical modification of lysine residues was to produce larger peptides than those obtained with a standard tryptic digestion in order to decrease sample background complexity without the use of the expensive Arg-C protease. After evaluating the efficiency of the two chemical reactions, we concluded that this strategy cannot be used to mimic an Arg-C digestion due to the non-quantitative character of the chemical reactions. Indeed during the modification process of lysine residues significant amounts of side products were generated which negatively impacted the complexity of the sample.

Chromatographic separation of non-tryptic digests

Enzymatic digests of biological samples are a very complex mixture of peptides. Before performing mass spectrometric analyses the peptide mixture is typically separated using reverse phase liquid chromatography. The separation is based on hydrophobic interactions of the peptides between the solid phase of the analytical column and its mobile phase. The separation of a tryptic digest is commonly performed on an analytical column containing silica beads chemically modified with octadecyl carbon chains (C_{18}). For non-tryptic digests (Lys-C or Arg-C) larger peptides are produced which are potentially more hydrophobic than the smaller tryptic peptides and may not elute from a C_{18} analytical column.

To estimate the hydrophobicity of the peptides generated by non-tryptic enzymes, the hydrophobicity index (SSRcalc hydrophobicity factor [108, 109]) of all peptides with a mass < 5 kDa but containing at least 5 amino acids obtained after the *in-silico* digestion of the human proteome with trypsin, Lys-C and Arg-C were calculated. The hydrophobicity index distribution is presented in Figure 40.

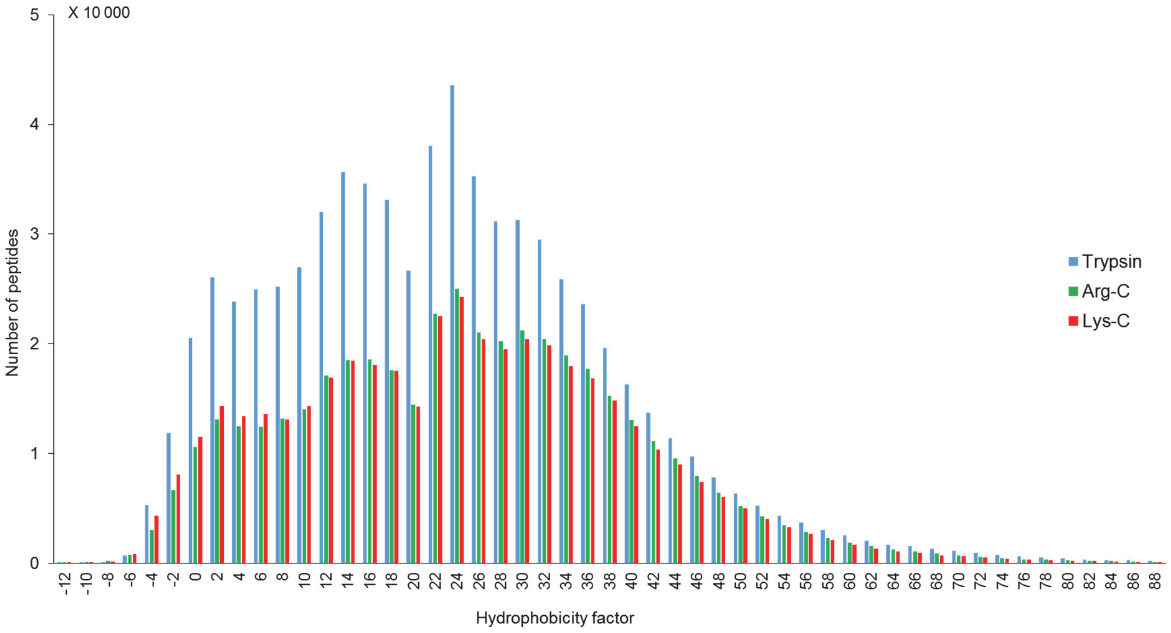


Figure 40: Theoretical distribution of the hydrophobicity index of peptides with a mass < 5 kDa but containing at least 5 amino acids obtained after the *in-silico* digestion of the human proteome with trypsin, Lys-C and Arg-C.

As previously observed and shown in figure 40, tryptic digestion generates a larger number of peptides in comparison with the Lys-C and Arg-C digestions. In the hydrophobicity range from 0 to 38 up to two times more peptides are produced with a trypsin digestion. Lys-C and Arg-C generate a similar number of peptides with a hydrophobicity index higher than 40 as compared to the standard tryptic digestion meaning that non-tryptic, Lys-C or Arg-C, digests do not require the use of a lower hydrophobicity analytical column such as C_8 or C_4.

LC-MS density of tryptic and Lys-C plasma digests

To experimentally confirm the decrease in background complexity observed with the *in-silico* digestions a depleted plasma sample was digested in parallel with trypsin and Lys-C. Both digests were analyzed by LC-MS under the same conditions *i.e.,* LC separation on a C_{18} analytical column and MS1 acquisition using a Q-Exactive plus HF mass spectrometer.

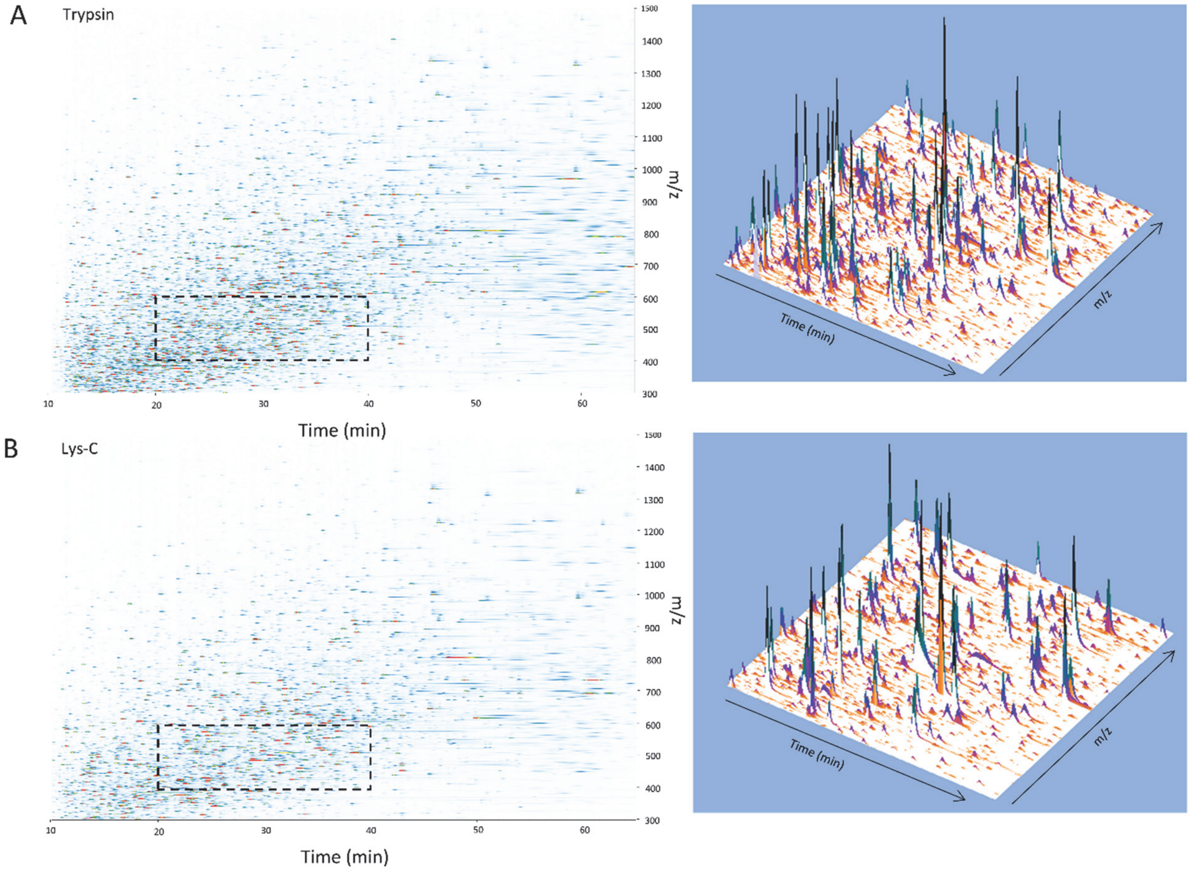


Figure 41: LC-MS ion map of a depleted human plasma digested with trypsin (A) and Lys-C (B). A 3 dimensional representation of the peak densities in the inserted, dashed rectangle is presented on the right panel.

Figure 41 presents the LC-MS ion maps (right) of the tryptic (A) and the Lys-C (B) digests displaying the intensity of each measured m/z from 300 to 1500 across the chromatographic separation from 10 to 65 minutes. As anticipated, the tryptic digestion shows a higher peptide density in comparison with the Lys-C digest especially between 10 and 50 minutes. The right panel displays a three dimensional representation of the zone between 20-40 minutes and 400-600 m/z, also illustrating the differential ion densities observed in the two digests. This decrease of ion density with Lys-C can be beneficial for targeted analyses due to a reduction of the number of ions which can be co-isolated with the precursor of interest thus limiting interferences.

Orthogonal application of proteolytic enzymes

The main interest of the use of alternative enzymes to trypsin is to improve sequence coverage. Lys-C and Arg-C are proteases which have a common cleavage site with trypsin, after lysine and arginine residues, respectively. Thus, Lys-C will be efficient to recover parts of proteins rich in arginine and Arg-C for those parts rich in lysine residues. Other enzymes cleave proteins at different sites such as Asp-N and Glu-C at the level of acidic amino acids that can potentially provide a higher degree of orthogonality than Lys-C or Arg-C to trypsin which can be beneficial to cover the unaccessible part of proteins by trypsin.

To evaluate the capabilities of these two types of alternative enzymes to access different parts of proteins within a proteome missed by trypsin, a simulation of the human proteome coverage for the tryptic, Lys-C and Glu-C digestions was performed by considering only peptides in the range of 8-25 amino acids.

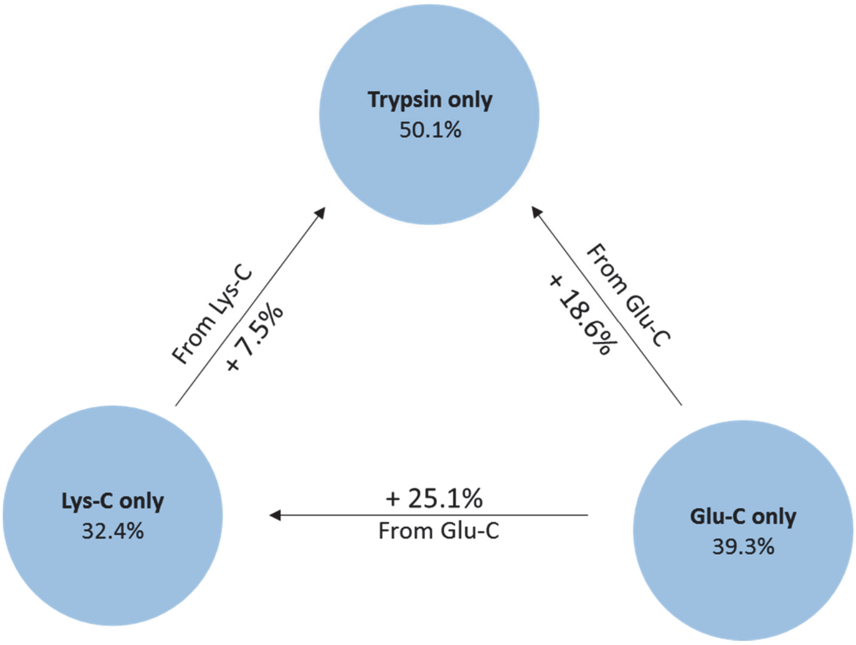


Figure 42: Theoretical amino acid coverage of the human proteome based on peptides in the 8-25 amino acid range.

In Figure 42 the simulated proteome coverage obtained for the three enzymes is presented. The simulation shows that the complementary use of Glu-C or Lys-C to the standard trypsin allows to identify an additional 18.6% or 7.5% of the sequences, respectively, as compared to a tryptic digestion. This result also indicates that trypsin and Glu-C have a higher degree of orthogonality than Lys-C and trypsin.

To experimentally assess the orthogonality of alternative enzymes to have access to the amino acid parts of proteins lost during tryptic digestions, a protein mixture (UPS1) containing 48 human proteins in equimolar amount was digested in parallel with trypsin, Lys-C and Glu-C before analysis by LC-MS/MS in DDA mode on a Q-Exactive Plus mass spectrometer. Peptide identification was performed by searching in a restricted database containing only the 48 proteins of interest. To determine the protein convergence for the three proteolysis experiments only peptides without any missed-cleavage and identified with a high confidence score (FDR < 0.01) were considered. Figure 43 summarizes the UPS1 protein coverage accessible to the different proteolysis reactions. As expected, the tryptic digestion obtained the highest protein coverage of 62.5% for UPS1 proteins.

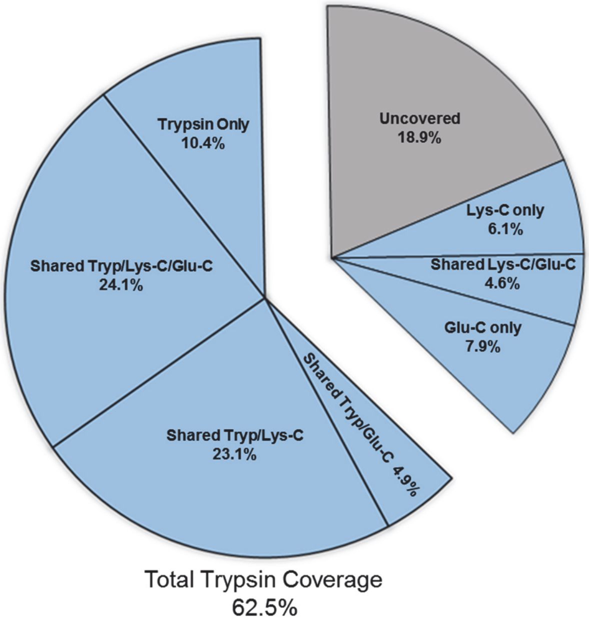


Figure 43: UPS1 experiment: amino acid protein coverage obtained by tryptic, Lys-C and Glu-C (considering only the cleavage after glutamic acid residues) digestions.

Lys-C and Glu-C (considering only the cleavage after glutamic acid residues) obtained lower protein coverages of 57.9% and 41.5% respectively which is in agreement with the simulation and other studies [44]. The lower proteome coverage obtained with Lys-C and Glu-C is mainly due to the search engines used for peptide identification which attribute lower ion scores to non-tryptic peptides because the algorithms are optimized for tryptic peptides. Using the identification information obtained with the three proteolysis reactions, the global UPS1 protein coverage reaches 81.1%. Glu-C and Lys-C allowed to identify 49.5% of the parts lost during tryptic digestion. The capability of alternative enzymes to disclose protein parts inaccessible to trypsin can be used to characterize mutated proteins or other protein isoforms. Glu-C was successfully used by Lesur et *al.,* to characterize a deletion mutation (amino acids 746-750) of the EGF receptor [118]. In this study, the standard tryptic digestion did not provide proteotypic peptides to characterize unambiguously the mutation due to the high density of lysine residues in the region of the deletion mutation.

3. Experimental selection of peptides to target eight lung cancer biomarker candidates

In the context of targeted bottom-up experiments for the quantification of low abundant proteins in a complex biological samples such as biomarkers in plasma, the selection of peptide targets is a crucial process which will drastically influence the reliability and the analytical performance of the assay. Before being used to quantify proteins, the surrogate peptides have to satisfy several criteria. First, they must be unique to the protein of interest (proteotypic peptides), second, they need to have good LC-MS physicochemical properties (amino acid composition, ionization, fragmentation and hydrophobicity) and third, they have to be efficiently produced during proteolysis. The empirical selection of peptide targets is a straightforward approach which is able to take into account all of the required criteria [129]. Here, the selection of the proteotypic peptides to target eight lung cancer protein biomarker candidates was based on the enzymatic digestion of recombinant proteins with five different enzymes trypsin, Lys-C, Arg-C, Asp-N and Glu-C.

Protein targets

In a recent article [22], the verification of 95 NSCLC biomarker candidates in clinical plasma samples from patients diagnosed with NSCLC and from apparently healthy individuals was reported using a highly multiplex targeted LC-SRM approach. As a result 17 proteins were identified as potential lung cancer plasma markers as they exhibited clear differential expression between healthy and diseased individuals. In the context of our project, eight of these potential protein lung cancer biomarkers (alpha-actinin-1, filamin-A, glucose-6-phosphate-1-dehydrogenase, endoplasmin, L-lactate dehydrogenase chain B, osteopontin, transaldolase and zyxin) were selected for the study of non-tryptic digests, based on their biological significance and their availability as recombinant proteins.

Selection of peptide targets for tryptic, Lys-C, Arg-C, Asp-N, and Glu-C digests

The selection of peptide targets for the eight proteins of interest, subjected to the five proteolytic conditions (trypsin, Lys-C, Arg-C, Asp-N and Glu-C), was empiric as described in Figure 44. First, the eight recombinant proteins were digested individually with trypsin, Lys-C, Arg-C, Asp-N and Glu-C. Second, the analysis of each digest was performed by LC-MS operating in DDA mode, and third, peptide identifications were performed by querying Mascot using a restricted database containing the sequences of the eight lung cancer biomarker candidates. Moreover, a manual curation was performed to determine if missing peptides were not present in the digest or if they were not identified by the search engine. To be a proper signature of a protein, peptides have to be specific *i.e.,* have a unique sequence. Each identified peptide was compared to the human protein database (UniProt) and only those peptides which were unique were retained.

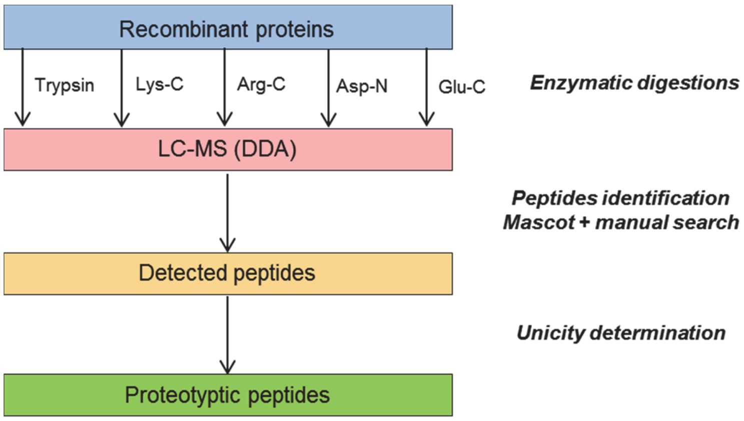


Figure 44*: Workflow for the empirical selection of proteotypic peptides to target eight proteins for five proteolysis experiments.*

To achieve a low LOD/LOQ in quantitative experiments the peptides should have a high ionization efficiency. To determine the best MS peptide responders for each of the proteins, generated by the five different digestion protocols, ion chromatograms of each peptide were generated for all the proteins under investigation using the extracted ion chromatogram (XIC). As example, the chromatograms of transaldolase digested with trypsin and Lys-C are presented in Figure 45. Using these data, identified proteotypic peptides were ranked by intensity for the different protein digests for each enzymatic condition in order to determine the peptide targets presenting the best response based on MS1.

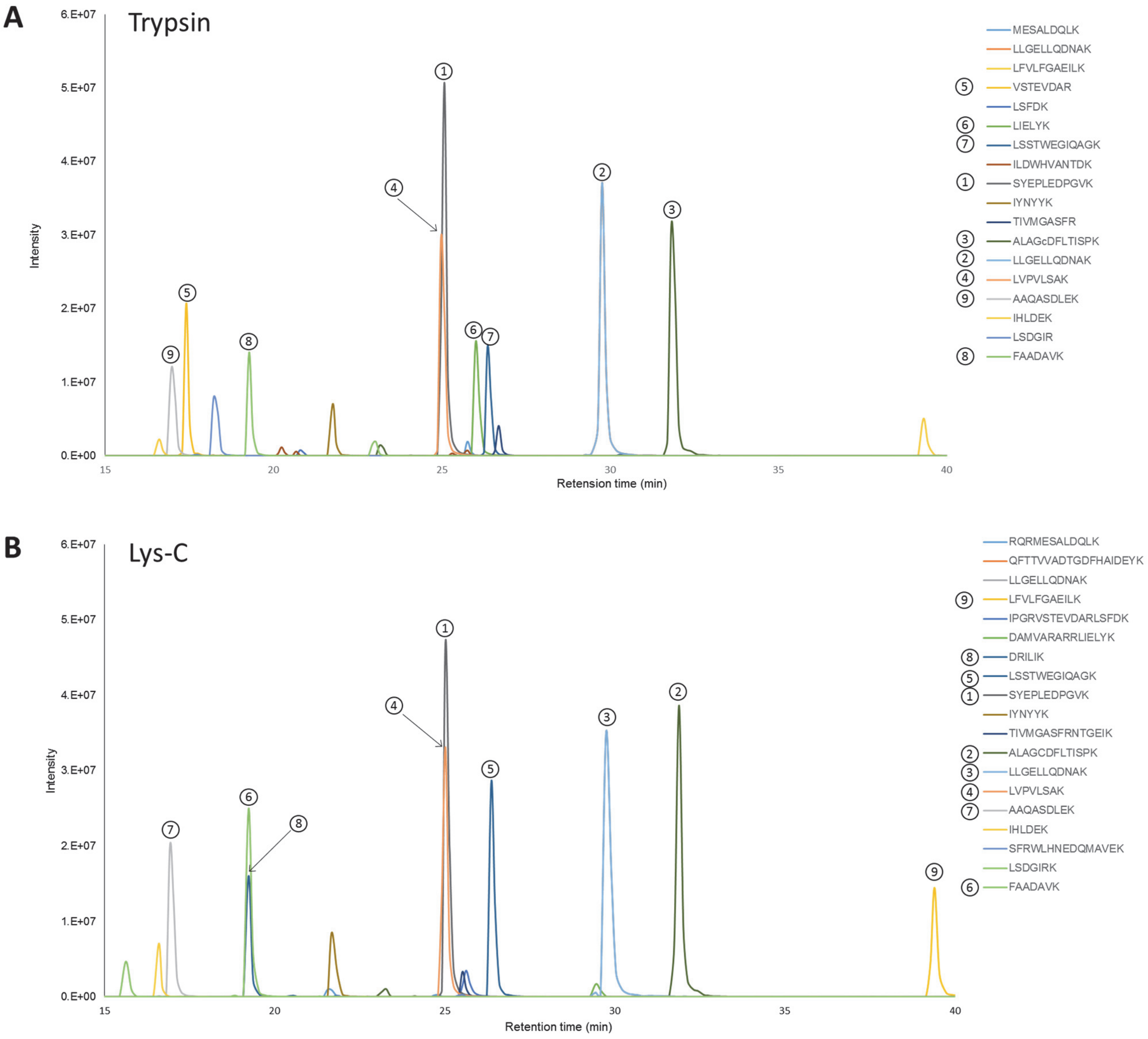


Figure 45: Extracted ion chromatograms of identified peptides in transaldolase tryptic (A) and Lys-C (B) digests.

Based on the acquired data, a total of 159 peptides were selected to target the eight proteins of interest in the five proteolytic conditions (trypsin, Lys-C, Arg-C, Asp-N and Glu-C).

4. Fragmentation of non-tryptic peptides

Depending on the enzymes used, different peptide sequences were generated due to the cleavage specificity of each protease (Table 9).

Digestion	Preceding AA	Peptide	Following AA	Types of peptide
Trypsin	XXXK	XXXXXXXXXR	XXXX	Trypsin only
	XXXR	XXXXXXXXXK	XXXX	Trypsin only
	XXXK	XXXXXXXXXK	XXXX	Commun trypsin/Lys-C
	XXXR	XXXXXXXXXR	XXXX	Commun trypsin/Arg-C
Lys-C	XXXK	XXXXRXXXXK	XXXX	Lys-C only
	XXXK	XXXXXXXXXK	XXXX	Commun trypsin/Lys-C
Arg-C	XXXR	XXXXKXXXXR	XXXX	Arg-C only
	XXXR	XXXXXXXXXR	XXXX	Commun trypsin/Arg-C
Asp-N	XXXX	DXXXXXXXXX	DXXX	Asp-N only
	XXXX	DXXXEXXXXX	DXXX	Asp-N including internal E
Glu-C	XXXE	XXXXXXXXXE	XXXX	Glu-C only
	XXXE	XXXXDXXXXE	XXXX	Glu-C including internal D

Table 9: Type of peptides generated by the five enzymes.

Trypsin, Lys-C and Arg-C are enzymes which specifically cleave proteins after basic amino acid residues *i.e.,* lysine and/or arginine. For these three proteolytic reactions the different peptides generated can be categorized in two groups. The first one includes peptides which are specific to a single enzymatic digestion, while the second one contains peptides which are shared between two different proteolytic reactions. In the case of trypsin, tryptic peptides which are specific to the tryptic digestion result from the cleavage of proteins after two different basic amino acids arginine and lysine. “Lys-C only” peptides are peptides which have a lysine as their C-terminal residue and contain internal arginine residues. “Arg-C only” peptides are peptides which have an arginine residue at their C-terminus and contain internal lysine residues. If protein cleavage by trypsin occurs between two identical basic amino acids, lysine and lysine or arginine and arginine, the generated peptides are considered as common Lys-C/trypsin and common Arg-C/trypsin peptides, respectively. Asp-N and Glu-C are proteases cleaving proteins near acidic amino acids, principally before aspartic acid and after glutamic acid residues, respectively. In this context, “Asp-N only” peptides are defined as peptides which are generated after the cleavage of proteins between two aspartic acid residues and “Glu-C only” peptides are defined as peptides which are generated after the cleavage of proteins between two glutamic acid residues. Asp-N and Glu-C do not have the same degree of specificity as trypsin. Indeed, these two enzymes can also cleave proteins before glutamic acid (Asp-N) and after aspartic acid (Glu-C) residues but to a lesser extent due to their slower reaction kinetics for these substrates.

The application of these enzymes results in the generation of non-tryptic peptides often containing internal basic amino acids which may influence their MS properties, such as ionization or fragmentation behaviors. In a previous study, Gallien et *al.* [72] presented the collision energy as a key parameter which drastically influences CID fragmentation. The optimal collision energy which generates the most intense fragments is a function of the peptide sequences, such as presence of easy cleavable amino acids like prolines [130, 131], the number of basic amino acids, or the charge state. The investigation of the effect of the collision energy on peptide fragmentation patterns was performed via the generation of pseudo-breakdown curves. Pseudo-breakdown curves show the evolution of the intensity of b- and y-fragment ions in an MS/MS spectrum acquired at different collision energies. Here, the pseudo-breakdown curves were generated by PRM analyses of 159 SIL peptides corresponding to the selected peptide signatures to target the eight NSCLC biomarker protein candidates under the five proteolytic conditions (trypsin, Lys-C, Arg-C, Asp-N and Glu-C). Tryptic peptides are normally short (10-16 amino acids), doubly charged peptides, for which two main fragmentation behaviors were observed. Peptide EEASDYLELDTIK (m/z 767.374, z = 2+) presented in Figure 46A illustrates a case where at low collision energy an abundance of low intensity fragments is produced. With an increase of the collision energy the fragment ion intensities reach a maximum at a normalized collision energy of (nCE) 20 (27.63 eV) before decreasing progressively at higher collision energies. In the second case, illustrated by the peptide AEAGVPAEFSIWTR (m/z 772.393, z = 2+) in Figure 46B, several distinct optimal collision energies are observed. At low collision energy two intense fragments are generated corresponding to the complementary b- and y- fragment ions formed by the cleavage of the proline residue. These two fragment ions reach a maximum intensity around nCE 15 (20.85ev) before decreasing progressively by undergoing a secondary dissociation at higher collision energies although another intensity maximum, albeit lower than the first one, was observed for a subset of fragment ions. The study of the pseudo-breakdown curves of Lys-C, Arg-C, Asp-N and Glu-C peptides showed similar fragmentation behaviors as observed for tryptic peptides. Indeed, as illustrated in Figure 46C and D the pseudo-breakdown curves for the peptides SILFVPTSAPRGLFDEYGSK (m/z 731.388, z = 3+) and ARVSSGYVPPPVATPFSSK (m/z 489.517, z = 3+) show, the presence of one or several optimal collision energies respectively similar as observed for tryptic peptides.

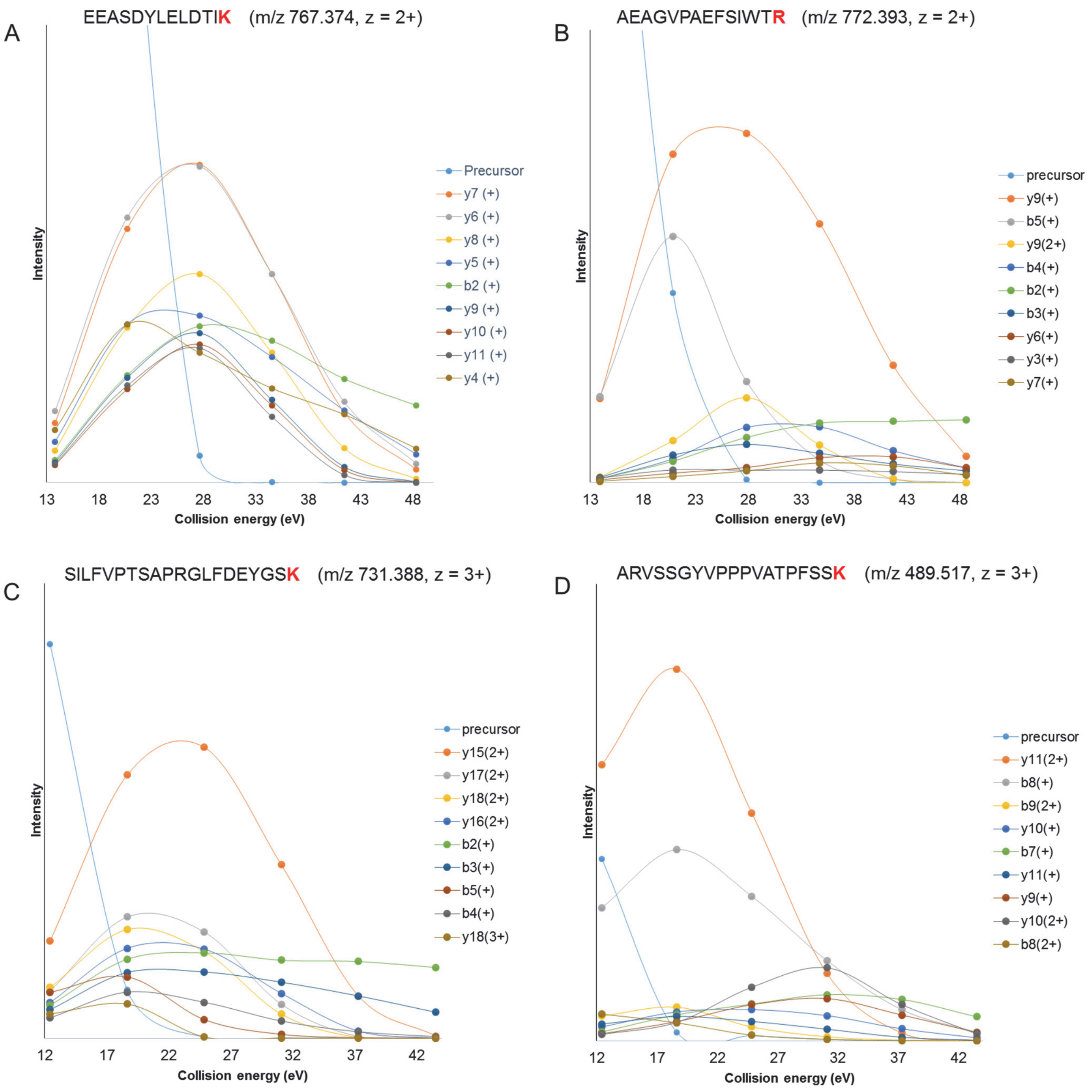


Figure 46: Pseudo-breakdown curves of EEASDYLELDTIK (m/z 767.374, z = 2+) (A), AEAGVPAEFSIWTR (m/z 772.393, z = 2+) (B), SILFVPTSAPRGLFDEYGSK (m/z 731.388, z = 3+) (C) and ARVSSGYVPPPVATPFSSK (m/z 489.517, z = 3+) (D) measured at six different nCE 10, 15, 20, 25, 30, 35. On the graphs nCE values were converted in eV using the equation: eV=nCEx(m/z / 500)xα where α is the charge factor,0.9 for z= +2 and 0.85 for z=3+.

Usually PRM experiments are performed by applying a unique value of “normalized” collision energy (nCE) for the entire set of targeted peptides. Derived from DDA experiments a default value of nCE from 25 to 30 has been widely used because it has been reported that the highest number of peptide identifications was obtained with these values using standard database searching algorithms [132-134]. In peptide identification studies it is more interesting to optimize the MS conditions to generate a wide fragmentation pattern in contrast with quantitative experiments for which the production of a few number of intense fragments would be beneficial. This consideration is illustrated in Figure 47. Figure 47A represents the pseudo-breakdown curve of AIPVAQDLNAPSDWDSRGK (m/z 683.348, z = 3+) including the evolution of the Mascot ion score for six collision energies. For this peptide at low collision energy (CE 17 ev), a small number of multicharged fragment ions is produced (intense y_{17}^{3+} and in smaller proportions y_{17}^{2+} and y_{15}^{2+}) as presented in Figure 47B. It has been identified with a low peptide Mascot ion score (22) due to the small number of assigned fragment ions. Indeed, to calculate the Mascot ion score many parameters are taken into account including the number of assigned singly and doubly charged fragment ions. By increasing the collision energy the number of assigned fragment ions is increased and consequently the peptide Mascot ion score rises progressively until it reaches a maximum (86) at 35eV despite the low intensity of the fragment ions (Figure 47C) .

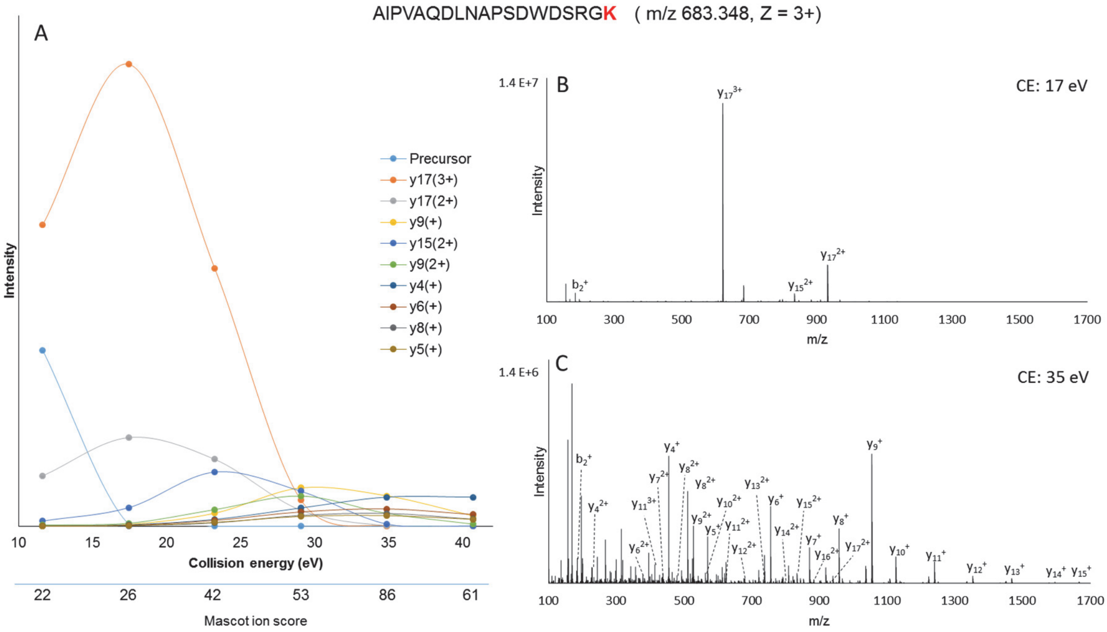


Figure 47: Pseudo-breakdown curve of AIPVAQDLNAPSDWDSRGK (m/z 683.348, z = 3+) measured at six different collision energies (nCE 10, 15, 20, 25, 30, 35) with the corresponding Mascot ion score (A). Annotated MS/MS spectrum acquired at CE 17eV and CE 35eV in panel B and C, respectively.

In PRM experiments the use of a default collision energy based only on m/z and charge state of the targeted peptides is far too restrictive to really reflect the specificities of the fragmentation of individual peptides. The sensitivity of PRM experiments benefits from a fine tuning of the collision energy, derived from the pseudo-breakdown curve of each peptide. To evaluate the gain induced by the adjustment of the collision energy, the 159 tryptic and non-tryptic synthetic peptides mentioned previously were used. For each peptide, the intensity of the most intense fragment ion across the 6 evaluated nCEs was compared to those measured at the “regular” 25 nCE. The results of this evaluation were grouped by peptide type and presented in Figure 48.

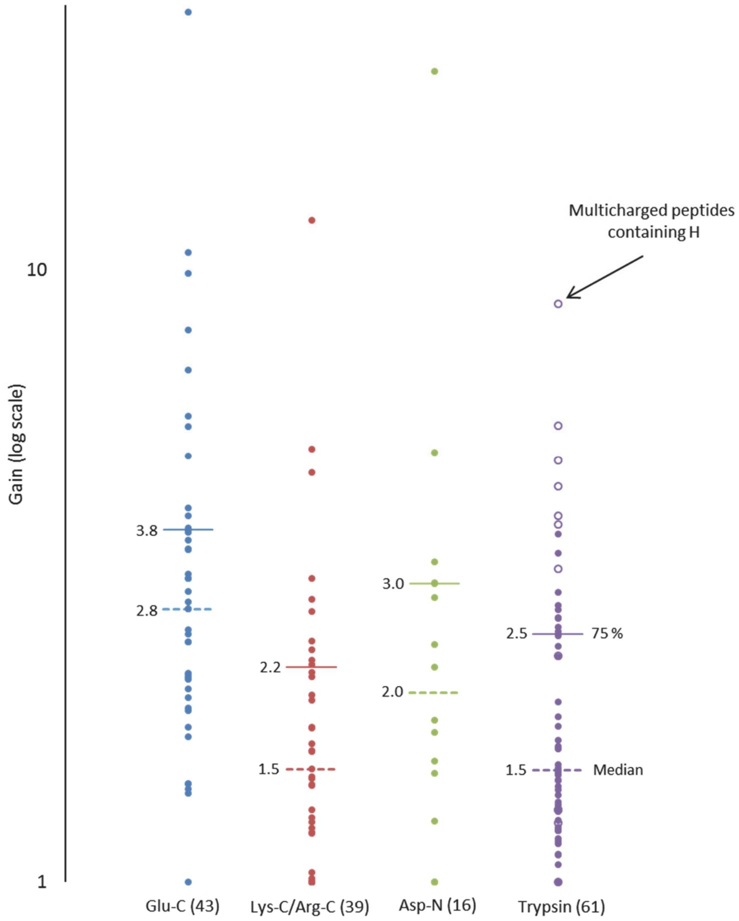


Figure 48: Gain in sensitivity (log scale) of 159 peptides categorized by peptide type based on the pseudo-breakdown curves of each peptide. The gain was defined as the ratio of the intensity of the most intense fragment ion across the 6 evaluated nCEs compared to that measured at the “regular” normalized collision energy of 25 nCE. Dashed lines represent the median value and solid lines represent the upper quartile value.

The gain in sensitivity presented in Figure 48 highlights clearly the benefits of peptide-specific optimization of the collision energy which allowed to increase the sensitivity of measurements up to 26-fold among our set of peptides. Peptide populations which show the highest median gain are Glu-C and Asp-N peptides with gains of 2.8 and 2.0, respectively. For tryptic and Lys-C/Arg-C peptides a 1.5 gain was observed. Inside the tryptic peptide population it is interesting to note that the six most intense gains observed corresponded to multicharged peptides containing internal histidines. Over all categories combined, a minimum gain of sensitivity of two fold was observed for more than a half of the peptides.

5. Conclusion

Currently trypsin is the most widely used enzyme in proteomics but is not always suitable for all types of proteomic studies. In this chapter it has been demonstrated that alternative enzymes should be considered to overcome the limitations of the standard proteomic workflow *i.e.,* the loss of sequence coverage and an increment in background complexity. Theoretical and experimental results showed that, in the context of targeted experiments, alternative enzymes allowed to recover parts of protein sequences lost during the trypsinization process, which is critical if these contain important information such as PTMs or amino acid mutations. The results also brought out that enzymes which have a single cleavage site such as Arg-C or Lys-C could have a significant impact on LC-MS density as it was illustrated on human plasma samples. Finally, it has been shown that a default collision energy, although yielding high identification scores during database searches, was not optimal for quantitative experiments. The sensitivity of targeted quantification assays in PRM mode can be improved by tuning the collision energy in order to produce a fewer number of intense fragment ions rather than a high number of fragments.

Chapter III: Accurate quantification in Lys-C digest

The methods developed and described in the previous sections were integrated and applied to the accurate quantification of lung cancer biomarker candidates in plasma samples. This chapter is the basis for a manuscript presently in preparation.

1. Introduction

In a recent bottom-up targeted proteomics study, Kim et *al.* identified several plasma proteins as potential NSCLC protein biomarkers [22]. These proteins exhibited differential expression in plasma from healthy individuals as compared to patients diagnosed with lung cancer. The further evaluation of such markers with the aim of translating them into a clinical assay required accurate quantification methods. Furthermore, robustness is also necessary to be able to share the quantitative results between laboratories.

The majority of quantitative assays involves a tryptic digestion while the use of this enzyme is not suitable to all studies, hence alternative enzymes may be required. To demonstrate the feasibility of an accurate protein quantification in complex biological samples digested with enzymes alternative to trypsin, we performed the accurate quantification of eight peptides, reported as potential biomarkers for NSCLC derived from alpha-actinin-1, filamin-A, transaldolase and zyxin. This was performed using PRM analyses of plasma samples from patients digested in parallel with Lys-C and trypsin and using concatenated polypeptides as internal standards.

2. Experimental design

To establish the proof-of-principle, the experiment was conducted on twenty four clinical plasma samples from twelve patients diagnosed with stage IV NSCLC and from twelve healthy individuals collected with the consent of each patient by filling the informed consent form approved by the Comité National d’Ethique de Recherche. The plasma samples were provided by the Integrated Biobank of Luxembourg (IBBL). The plasma samples from cancer patients and apparently healthy individuals were generated after the centrifugation of the blood samples collected in several Luxembourgish hospitals from men aged between 54 and 69 years, who were smokers. Each plasma sample was processed as described in Figure 49.

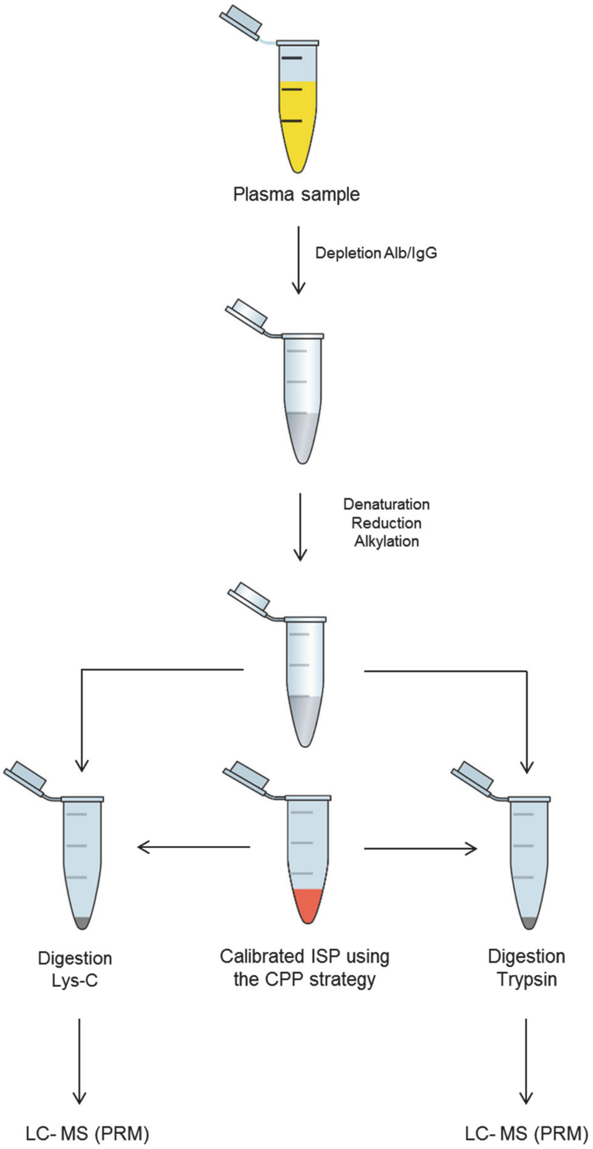


Figure 49: Plasma sample preparation for an accurate quantification of eight peptide targets in two different enzymatic digests.

Four candidate NSCLC biomarkers, alpha-actinin-1, filamin-A, transaldolase, and zyxin, were quantified using the concatenated polypeptide method combined with non-tryptic proteolysis (Lys-C digestion). To decrease the dynamic range of protein concentration, the two most abundant proteins, *i.e.*, albumin and immunoglobulins, were removed. After denaturation, reduction of the disulfide bonds and alkylation of the free thiol groups, each sample was split into two equal parts which were digested with either Lys-C or trypsin. Before the LC-MS/MS analysis, each digest was supplemented with a defined amount of eight isotopically-labeled signature peptides representing the four proteins of interest under the two proteolytic digestion conditions. The internal standards were calibrated beforehand using the concatenated polypeptide strategy. In this experiment, two peptide surrogates of each protein and generated by both Lys-C and tryptic digestions were considered.

Design of concatenated polypeptides

As stated in chapter I, the selection of the labeling scheme of the reporter peptides requires special attention to maximize the number of isotopically-labeled signature peptides that can be concomitantly calibrated in a single analysis. It has to be noted that individual measurements can make use of the same reporter, but a biomarker panel requires the design of a unique reporter combination. Following the methodology described in chapter I, the design of the polypeptides corresponding to the eight targeted peptides was made in order to be able to calibrate sets of four isotopically-labeled signature peptide using a four isotopologue calibration mixture (AALPAAFK, AALPAAFK, AALPAAFK and AALPAAFK). The two sets of four concatenated polypeptides defined are presented in Table 10. Each polypeptide was trypsinized and the different digests were pooled together according to the two sets defined.

Table 10: Sequences and labeling scheme of the two sets of concatenated polypeptides designed.

Set 1	Signature peptide	Reporter
Zyxin	FSPGAPGGSGSQPNQK	AALPAAFK
	FSPVTPK	AALPAAFK
Filamin-A	SPFSVAVSPSLDLSK	AALPAAFK
	DAGEGGLSLAIEGPSK	AALPAAFK

Set 2	Signature peptide	Reporter
Alpha-actinin-1	LVSIGAEEIVDGNVK	AALPAAFK
	DDPLTNLNTAFDVAEK	AALPAAFK
Transaldolase	LVPVLSAK	AALPAAFK
	SYEPLEDPGVK	AALPAAFK

Optimization of the acquisition parameters

Prior to the calibration of the eight isotopically-labeled signature peptides used as internal standards, the optimal collision energy for each peptide has to be determined. Usually, in quantitative experiments, four to eight fragments are considered to confirm the identity of the measured peptides and to assess the quality of the acquired MS/MS spectra. In the absence of interferences, the sensitivity of measurement is limited by the selected fragment with the lowest intensity. In this study, peptide quantification was performed using the four most intense fragment ions (the minimum number required) to obtain the optimum sensitivity while keeping maximum selectivity. The optimal collision energy was thus defined as the one for which the intensity of the fourth most intense fragment achieves a maximum to ensure enough selectivity to confirm the peptide intensity. As described in chapter II, the breakdown curves of the eight peptides were plotted in order to evaluate the variations of peptide fragment intensities (y- and b-ion series) at six different collision energies (nCE 10, 15, 20, 25, 30 and 35). The determined optimal collision energies for each peptide are labeled on the pseudo-breakdown curves displayed in Figure 50.

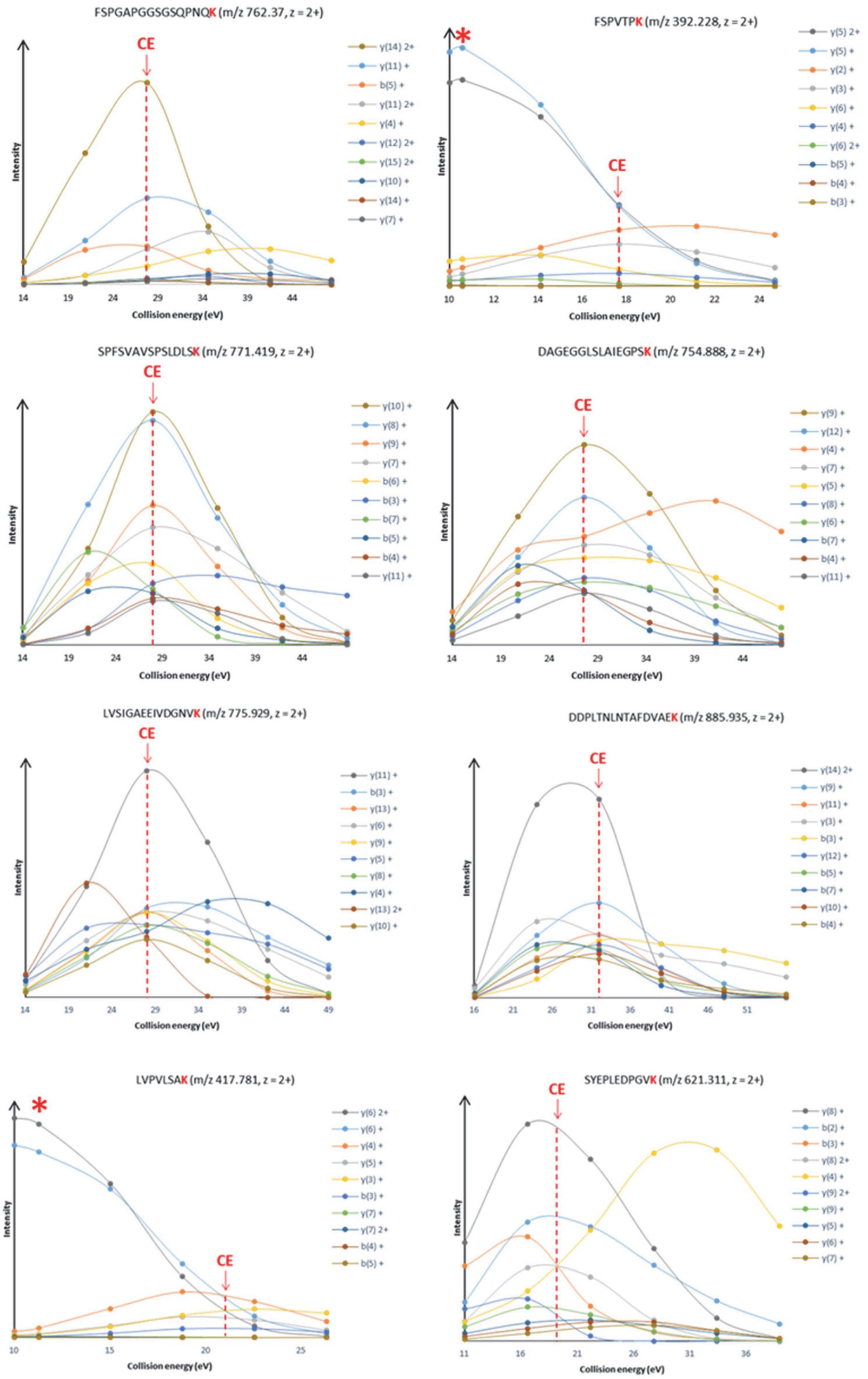


Figure 50: Pseudo-breakdown curves of ten precursors measured at six different collision energies (nCE 10, 15, 20, 25, 30, 35) of eight ISPs. The optimal collision energy of each peptide is indicated with red line arrows. The red star indicates the collision energy without taking into account the need for selectivity.

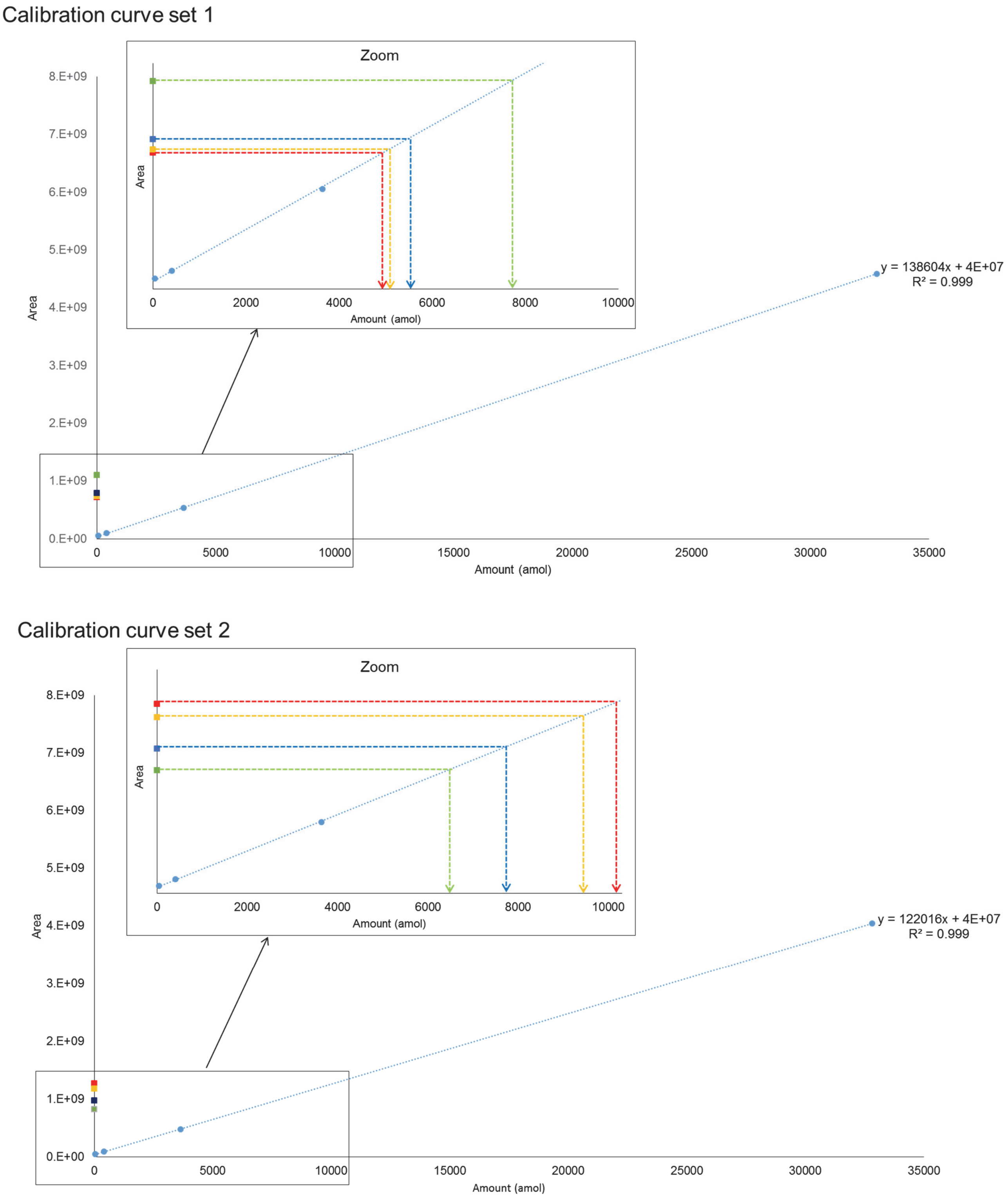
Calibration curves of signature peptides

In the calibration process of the eight isotopically-labeled standards, the two sets of concatenated polypeptide digests were supplemented with a defined amount of calibrated reporter isotopologues. The analyses were carried out in triplicate in PRM mode on a Q-Exactive Plus mass spectrometer. During the analyses of the two sets, the eight reporters and the four isotopically-labeled signature peptides were monitored at their optimized collision energy. As shown in Figure 51, the two calibration curves were built using the area under the curve (AUC) determined from the fragment traces of the four isotopologue reference reporters. They were used to accurately quantify the cleavable reporter peptides, and consequently the corresponding equimolar isotopically-labeled signature peptides.

The PRM analysis of concatenated polypeptide digests also yields information on the signature peptides *i.e.,* the full MS/MS spectra used as reference for quantification. The partial MS/MS spectra of the eight isotopically-labeled signature peptides are presented in Figure 52. They contain only the four fragment ions of interest extracted from the full MS/MS spectra acquired at their optimal collision energy.

Analyses of clinical samples

Following their calibration and the determination of their reference spectra, the eight isotopically-labeled signature peptide standards were spiked in a defined amount into all plasma samples. The PRM acquisition of the eight pairs of endogenous peptides/calibrated internal standards was performed in analytical triplicates on a Q-Exactive Plus mass spectrometer.



Calibration curve set 1

Calibration curve set 2

Figure 51: Calibration of the eight internal standards distributed into groups using isotopologue calibration curves

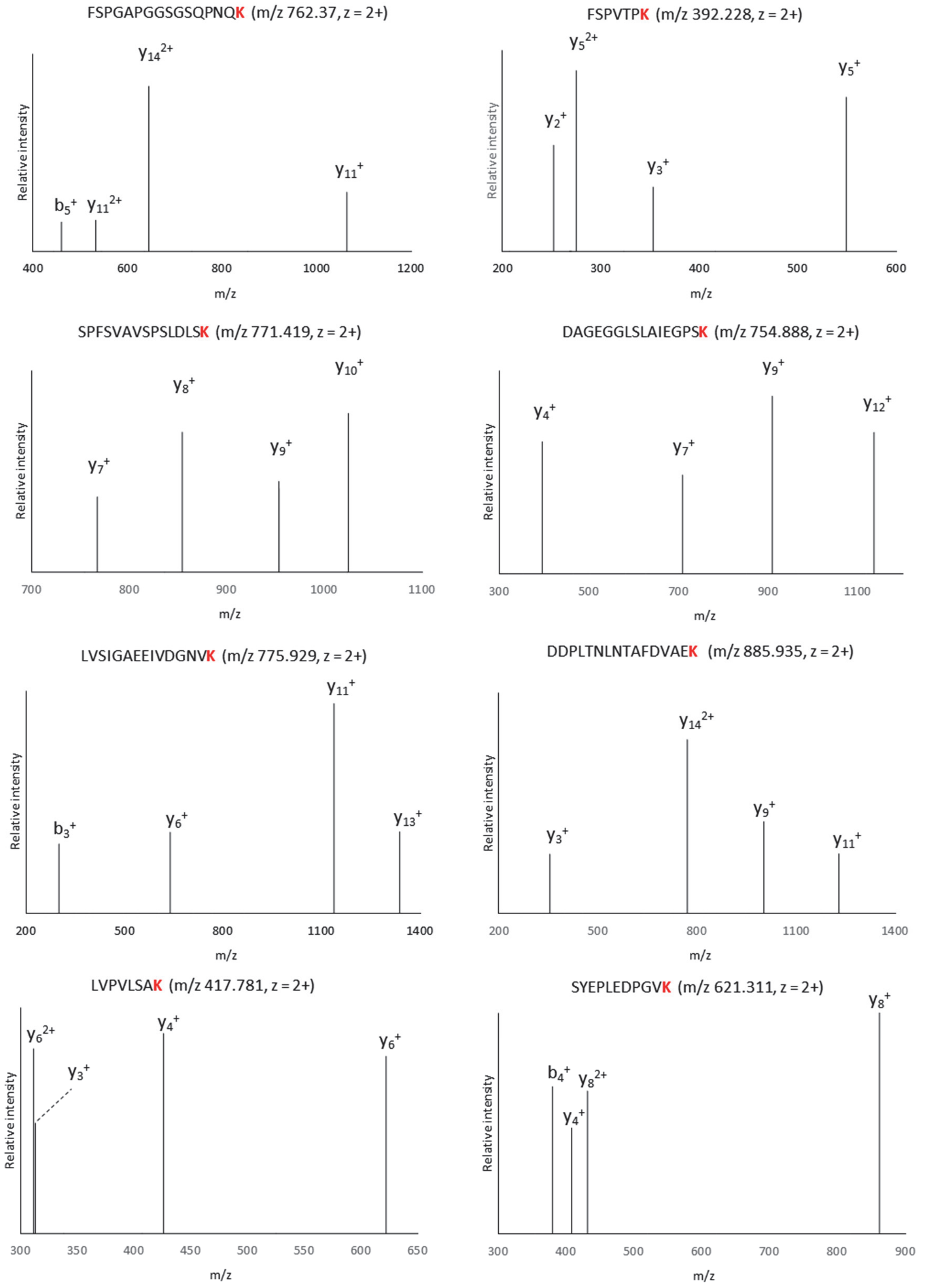


Figure 52: Partial MS/MS spectra of the eight peptide targets, showing the four fragments of interest per peptide.

3. Data processing and results

The quantitative PRM data were processed in two steps. First, the peptide identity was confirmed based on the matching of the acquired full MS/MS spectrum with a reference spectrum. Second, the quantification was performed using the traces of fragment ions of confirmed peptides which presented good purity *i.e.,* which did not present interferences (Figure 53).

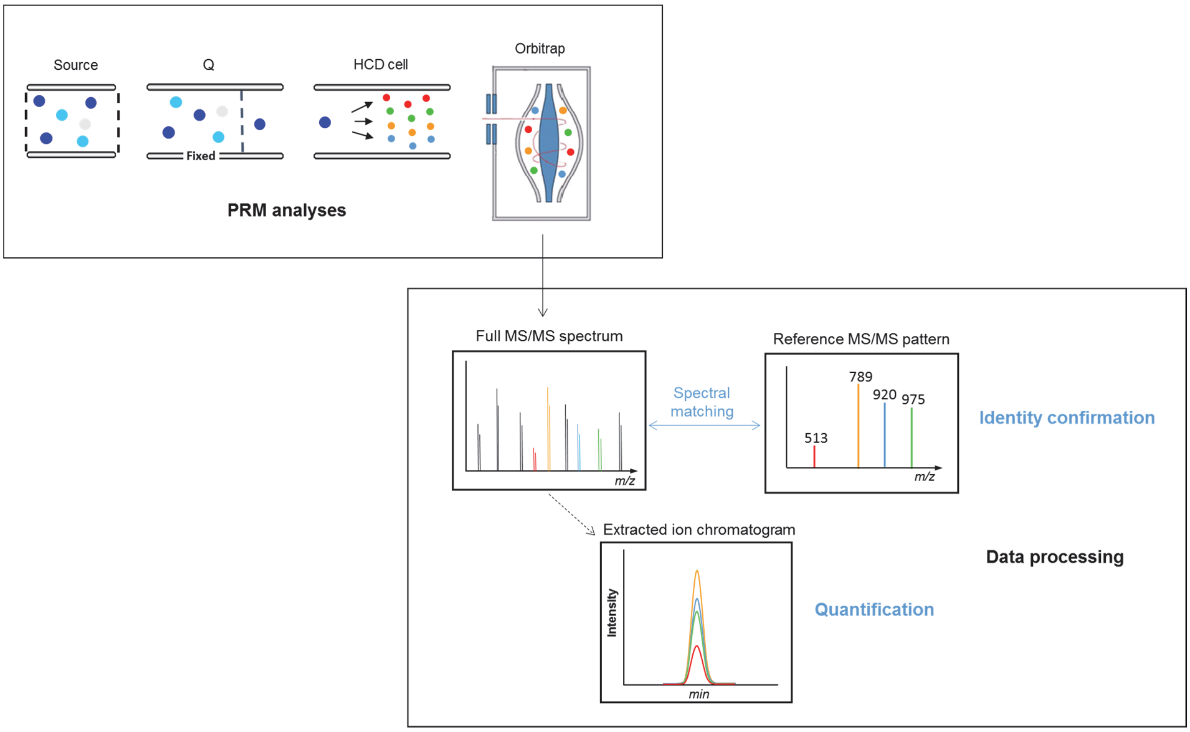


Figure 53: PRM data processing workflow (figure adapted from Gallien et *al.*, Methods, 2015: 15: 15-23).

Confirmation of peptide identity and evaluation of interferences

To confirm the peptide identity, the fragmentation patterns of the peptide were compared to their reference MS/MS spectra acquired in buffer. The similarity evaluation was based on the comparison of the relative intensities of fragments through the calculation of a spectral contrast angle θ [135]. In this process, the MS/MS spectrum of the peptide to evaluate and its reference spectrum are considered as two vectors which have as coordinates the intensities of their selected fragment ions. The angle θ was defined as presented in equation 1 and was used to estimate the similarity between the two MS/MS spectra.

$$\cos\theta = \frac{\sum_{i=1}^{n} I_{exp_i} \times I_{ref_i}}{\sqrt{\sum_{i=i}^{n}(I_{exp_i})^2} \times \sqrt{\sum_{i=i}^{n}(I_{ref_i})^2}} \qquad \text{(Equation 1)}$$

In Equation 1, $I_{exp_i}$is the intensity of the fragment *i* in the tandem mass spectrometry spectrum under assessment and $I_{ref_i}$is the intensity of the fragment *i* in reference MS/MS spectrum. Figure 54 is a two dimensional illustration of the use of the spectral contrast angle to measure similarities between two MS/MS spectra. Only two fragments were taken into account for simplicity of the illustration. In Figure 54A, the fragmentation pattern of the peptide analyzed (red) is significantly different from the reference MS/MS spectrum (blue). It results in a relative high spectral contrast angle value (23°) between the two vectors, which have different directions. In contrast, in Figure 54B, the two MS/MS spectra have perfectly identical fragmentation patterns with different abundances so in this context the corresponding vectors are collinear with different lengths. The identity of the peptide is considered as confirmed if the MS/MS spectrum of the peptide under evaluation presents high similarity with the reference spectrum i.e. the spectral contrast angle is below 12° (cos θ > 0.98) [136, 137].

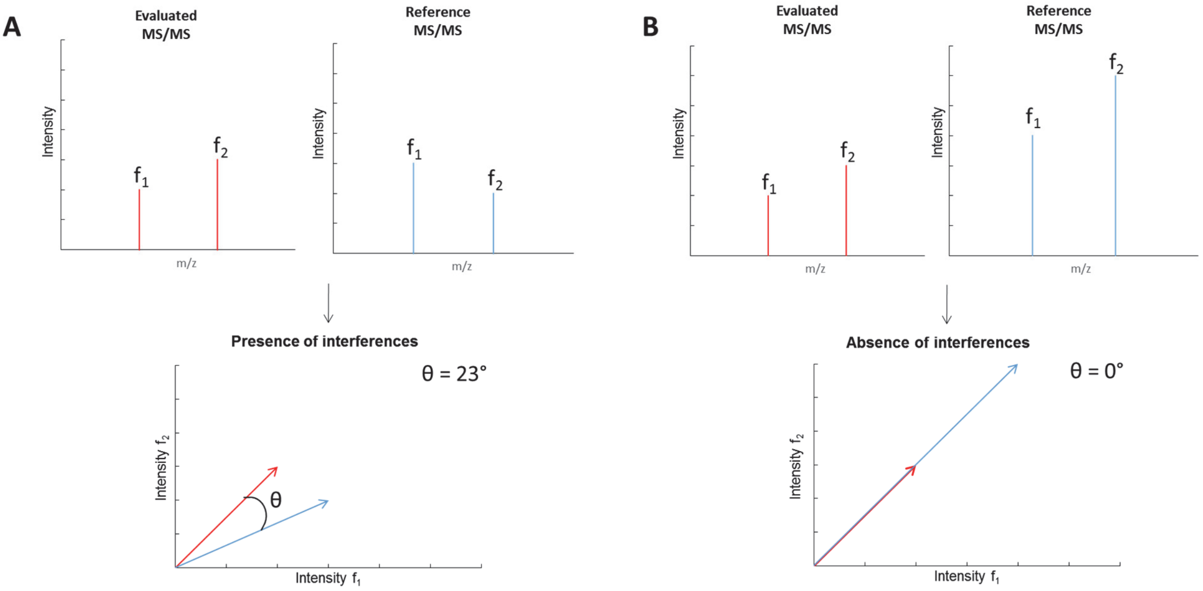


Figure 54: Representation of the spectral contrast angle θ, used to measure spectral similarities (figure adapted from Katty et *al.* Journal of The American Society for Mass Spectrometry, 2002; 1: 85-8).

The spectral contrast angle using the four most intense fragment ions of the eight endogenous peptides and their corresponding ISPs was calculated. Three cases can be observed. In the first case, the spectral contrast angle is below 12°, indicating that no interferences are observed for the evaluated fragments allowing for quantification (Figure 55A). In the second case, the determined spectral contrast angle is higher than 12°, indicating that at least one fragment shows interference or poor S/N (peptide in very low abundancy) (Figure 55B). To determine which fragment ion induced distortion in the fragmentation pattern the deviations of fragment ion intensities between the evaluated and the reference MS/MS spectrum were calculated. If only one fragment presents discrepancies (deviation in fragment intensity higher than 40%) it can be replaced with another fragment and the spectral contrast angle was recalculated. If θ drops below 12°, the targeted peptide can be quantified (Figure 55C). The last case corresponds to the situation where the spectral contrast angle remains higher than 12° in spite of the fragment ion substitution. In this last case, the peptides cannot be measured reliably and are excluded from the quantification process.

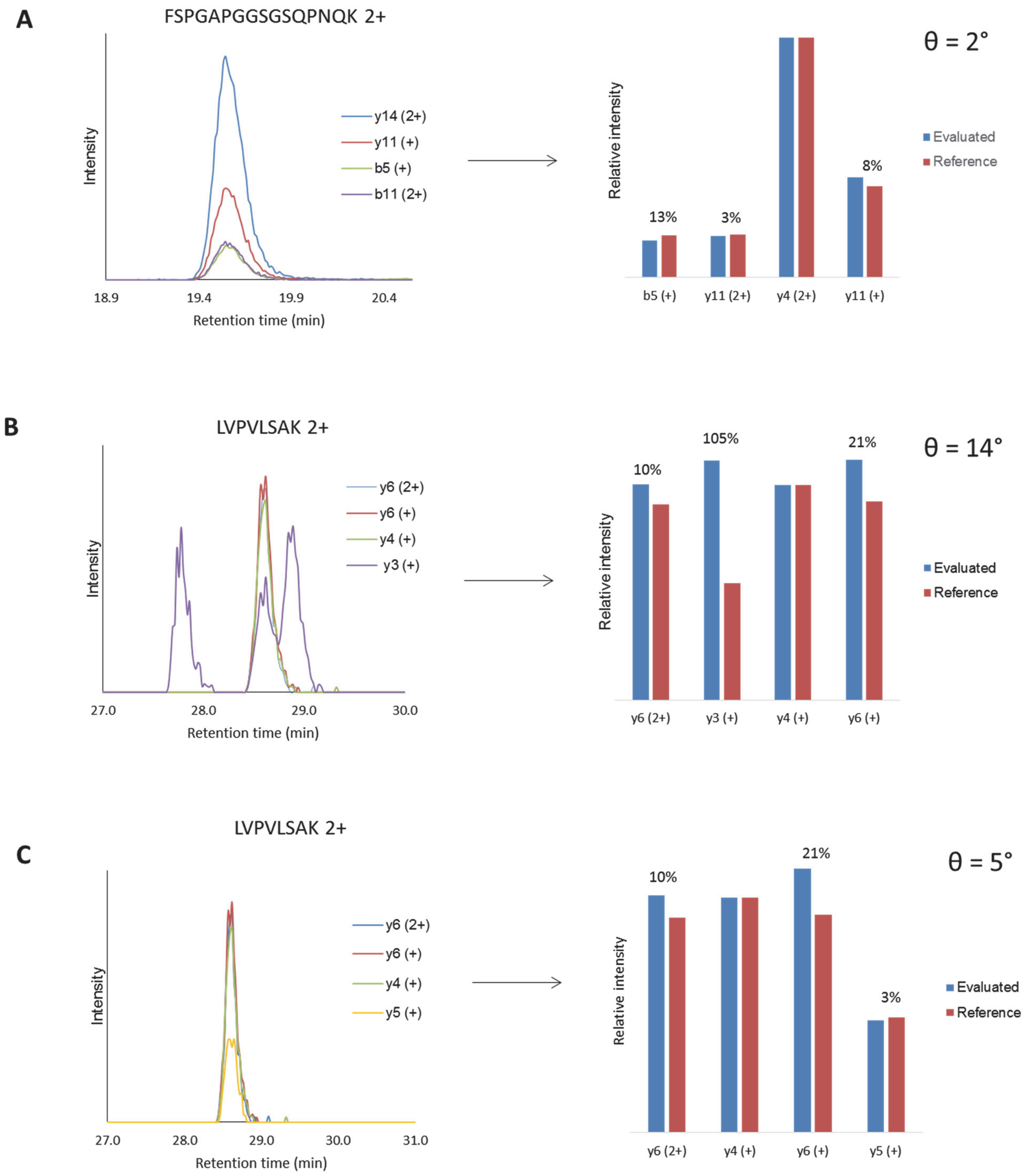


Figure 55: Illustration of the use of the spectral contrast angle and deviations in fragment ion intensities (%) to evaluate the signal of experimental data. Peptide FSPGAPGGSGSQPNQK can be quantified θ = 2° (A), peptide LVPVLSAK cannot be quantified θ = 14° (B), peptide LVPVLSAK after exchange of the fragment y_3^+ (deviation in fragment intensity of 105%) with y_5^+ can be quantified θ =5° (C).

Accurate quantification of endogenous peptides

The accurate quantification of the surrogate peptides of the four lung cancer biomarkers in digested plasma samples was carried out based on the most intense fragments free of interferences of the endogenous and the internal standard peptides.

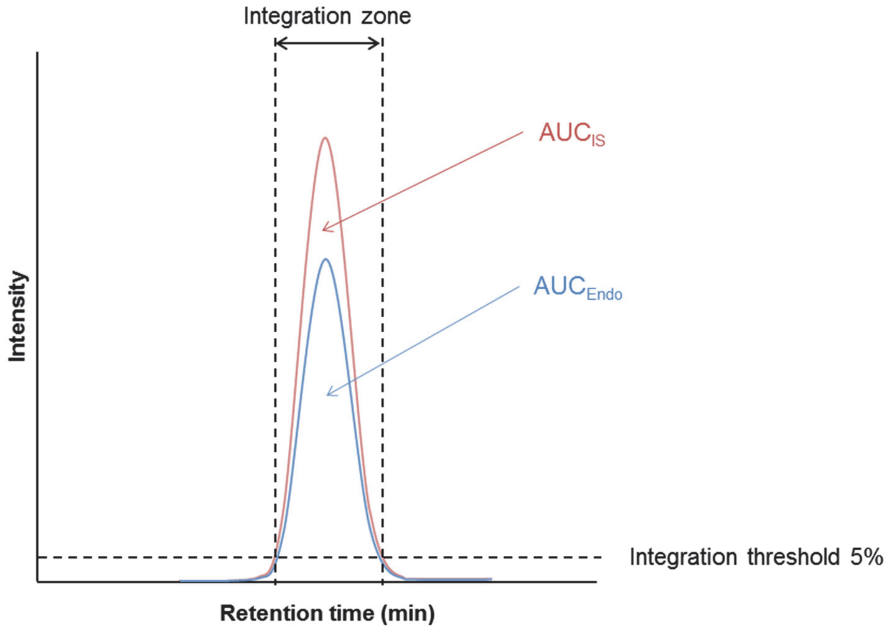


Figure 56: Extracted ion chromatograms of the most intense fragments for the endogenous (blue) and the internal standard (red). The areas under the curve (AUC) are defined for the integration zone.

As illustrated in Figure 56, the signal of the most intense fragment of the endogenous peptide and the internal standard were extracted from the PRM data set analyses with a 10ppm tolerance. The quantification of the endogenous peptides was performed by comparing the areas under the curve (AUC) of the integrated traces (one point calibration) defined using an intensity threshold of 5%. The accurate amount of endogenous peptide was determined using Equation 2:

$$\mathrm{Amount}_{\mathrm{endo}} = \frac{\mathrm{AUC}_{\mathrm{endo}}}{\mathrm{AUC}_{\mathrm{IS}}} \times \mathrm{Amount}_{\mathrm{IS}} \qquad \text{(Equation 2)}$$

Using this equation, the concentrations of the eight endogenous peptides in the 24 plasma samples generated by the two different proteolysis were determined by applying the criteria previously described. An overview of the quantitative results is presented in Figure 57.

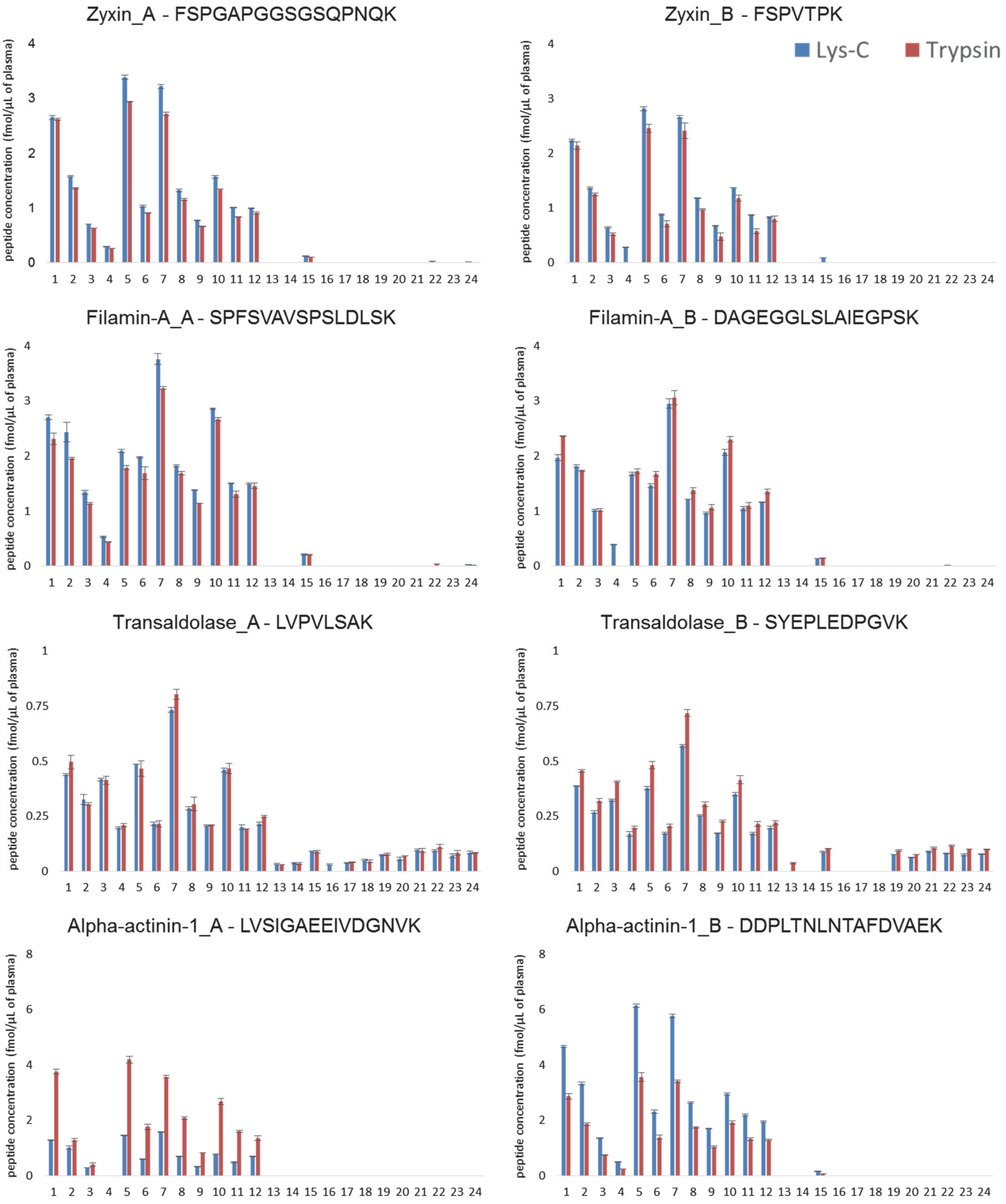


Figure 57: Determination of the accurate concentration of eight surrogate peptides of four lung cancer biomarkers; Lys-C (blue) and tryptic (red) plasma digests. Patients diagnosed with NSCLC (n°1 to 12) and healthy individuals (n°13 to 24).

Each histogram presents the determined concentrations of one endogenous peptide in the twenty four plasma samples digested by Lys-C and trypsin in blue and red, respectively. Regardless of the enzyme used, differential expression of all peptides was observed between the plasma of diseased individuals and controls. The peptides were successfully quantified in most plasma samples from patients diagnosed with lung cancer, while they were only quantifiable in a few control samples due to their low abundance (below the limit of detection). This accurate quantitative experiment was performed only on stage IV NSCLC plasma samples. It confirmed the results previously obtained on a large cohort, involving 72 patients diagnosed with NSCLC stage I to IV and 30 healthy volunteers, which highlighted alpha-actinin-1, filamin-A, transaldolase and zyxin as promising lung tumor markers [22].

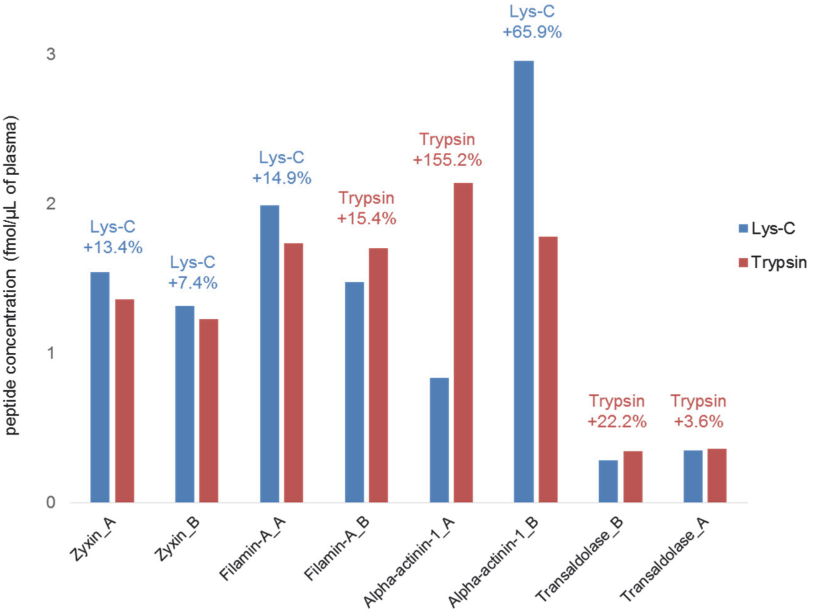


Figure 58: Average peptide concentrations determined in the 12 plasma samples from patients diagnosed with NSCLC.

Figure 58 displays the average peptide concentration determined in the twelve plasma samples from patients diagnosed with NSCLC for the two proteolytic digestions. On average, for Zyxin_A (FSPGAPGGSGSQPNQK), Zyxin_B (FSPVTPK), Filamin-A_A (SPFSVAVSPSLDLSK) and Alpha-actinin-1_A (LVSIGAEEIVDGNVK), the concentrations measured for the corresponding proteins were around 13%, 7%, 15% and 66% respectively higher using a Lys-C digestion as compared to trypsin. Conversely, the measured concentrations for Filamin-A_B(DAGEGGLSLAIEGPSK), Alpha-actinin-1_B(DDPLTNLNTAFDVAEK), Transaldolase_A (LVPVLSAK) and Transaldolase_B (SYEPLEDPGVK) were around 15%, 155%, 22% and 4%

respectively higher using a tryptic digestion. Regarding the comparison between the Lys-C and the tryptic digestions, a similar performance was observed for peptide quantification using six peptides, Zyxin_A, Zyxin_B, Filamin-A_A, Filamin-A_B, Transaldolase_A and Transaldolase_B. For the two other peptides, Alpha-actinin-1_A and Alpha-actinin-1_B, more significant discrepancies were observed between the two proteolyses.

These variations between the Lys-C and tryptic digestions reflect differences in the proteolytic efficiency of the two enzymes. It has been reported that the amino acid environment close to the cleavage sites has an influence on proteolytic efficiency [138]. In this study, the difference between the two proteolytic results were small for the large majority of peptides except for the peptides DDPLTNLNTAFDVAEK and LVSIGAEEIVDGNVK. For DDPLTNLNTAFDVAEK, the Lys-C digestion was more efficient than the tryptic one, likely due to the presence of an arginine residue near the N-terminal cleavage site (---KLRKDDPLTNLNTAFDVAEKMLDA---), which inhibits trypsin activity, whereas it does not affect Lys-C. In the tryptic digest a large quantity of KDDPLTNLNTAFDVAEK was observed. Regarding the generation of the peptide LVSIGAEEIVDGNVK a lower efficiency can be hypothesized for both digestions based on the concentration determined for the other Lys-C generated peptide of the same protein (DDPLTNLNTAFDVAEK). While the prediction of missed cleavage sites for trypsin has been studied for decades, the literature dedicated to missed cleavages by Lys-C is far less abundant. Similar to trypsin, the Lys-C digestion can be completely or partially affected by the presence of amino acids such as glutamic and aspartic acids or lysines close to the cleavage sites. In the present case (---KGVKLVSIGAEEIVDGNVKMTLG---) these reasons cannot explain the lower digestion efficiency noticed for the generation of this peptide in both digestions, and especially for Lys-C proteolysis. This experiment demonstrates that a quantitative assay involving a Lys-C digestion can achieve in most cases similar or higher performance as compared to the standard tryptic digestion. Its use is thus particularly indicated for studies requiring alternative enzymes, for instance to target part of proteins inaccessible to trypsin.

In bottom-up approaches, to generate consistent data, the proteolysis has to be reproducible across the entire sample set. Trypsin presents high reproducibility capabilities, which explain its general usage. To evaluate if Lys-C exhibits a similar performance, the ratio α as defined in Equation 3 was calculated for each peptide and each plasma sample (n°1 to 12).

$$\alpha = \frac{[Peptide\ x]_{Lys-C\ digest}}{[Peptide\ x]_{Tryptic\ digest}} \qquad \text{(Equation 3)}$$

α represents the ratio of peptide concentration determined for one peptide in the Lys-C digestion divided by the one determined in the tryptic proteolysis. A constant value for α through a large sample set indicates consistency between the two proteolysis. The evolution of this ratio across the twelve plasma samples from patients diagnosed with NSCLC was plotted in Figure 59 for each targeted peptide.

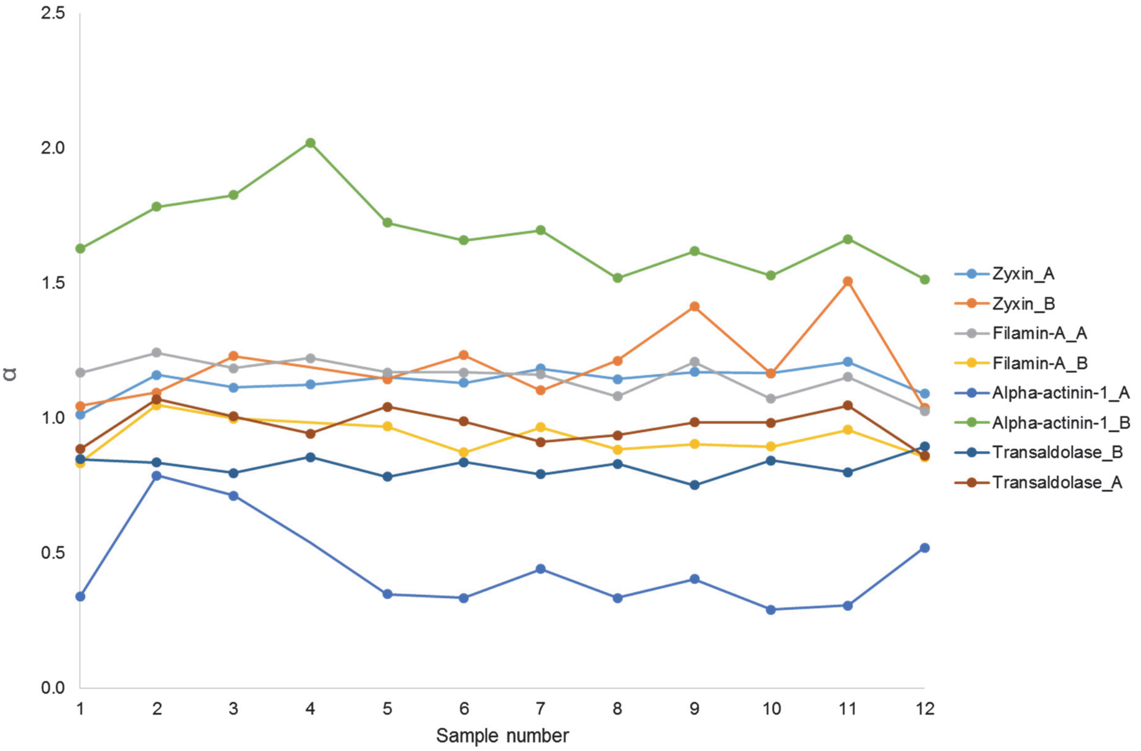


Figure 59: Variation of α across the twelve plasma samples from patients diagnosed with NSCLC.

In general, trypsin and Lys-C digestions show high consistency for six peptides including zyxin_A, filamin-A_A, filamin-A_B, alpha-actinin-1_B, transaldolase_B, transaldolase_A (variations of α <8%). Regarding zyxin_B, a good consistency between both digestions was observed as illustrated by the 11% variation of α. Only alpha-actinin-1_A, exhibited significant variations in the ratio of concentrations (α of 37 %).

In this experiment, each of the four proteins of interest was quantified through two peptides located in different parts of the protein. Regardless of the enzyme used, the quantitative values determined for the two peptides were consistent (Figure 57), meaning that protein concentration can be determined using both peptides. Bottom-up quantitative proteomics relies on the equimolarity between the released peptides and the native endogenous protein, which allows the molar concentrations of the endogenous peptides in plasma to be considered as equal to the molar concentration of proteins in plasma. To determine the mass concentration of the targeted proteins in plasma, the molecular concentrations were multiplied by the molecular mass of the corresponding protein. Figure 60 reports the mass concentration of the proteins in each clinical sample using the eight peptides for the two digestions.

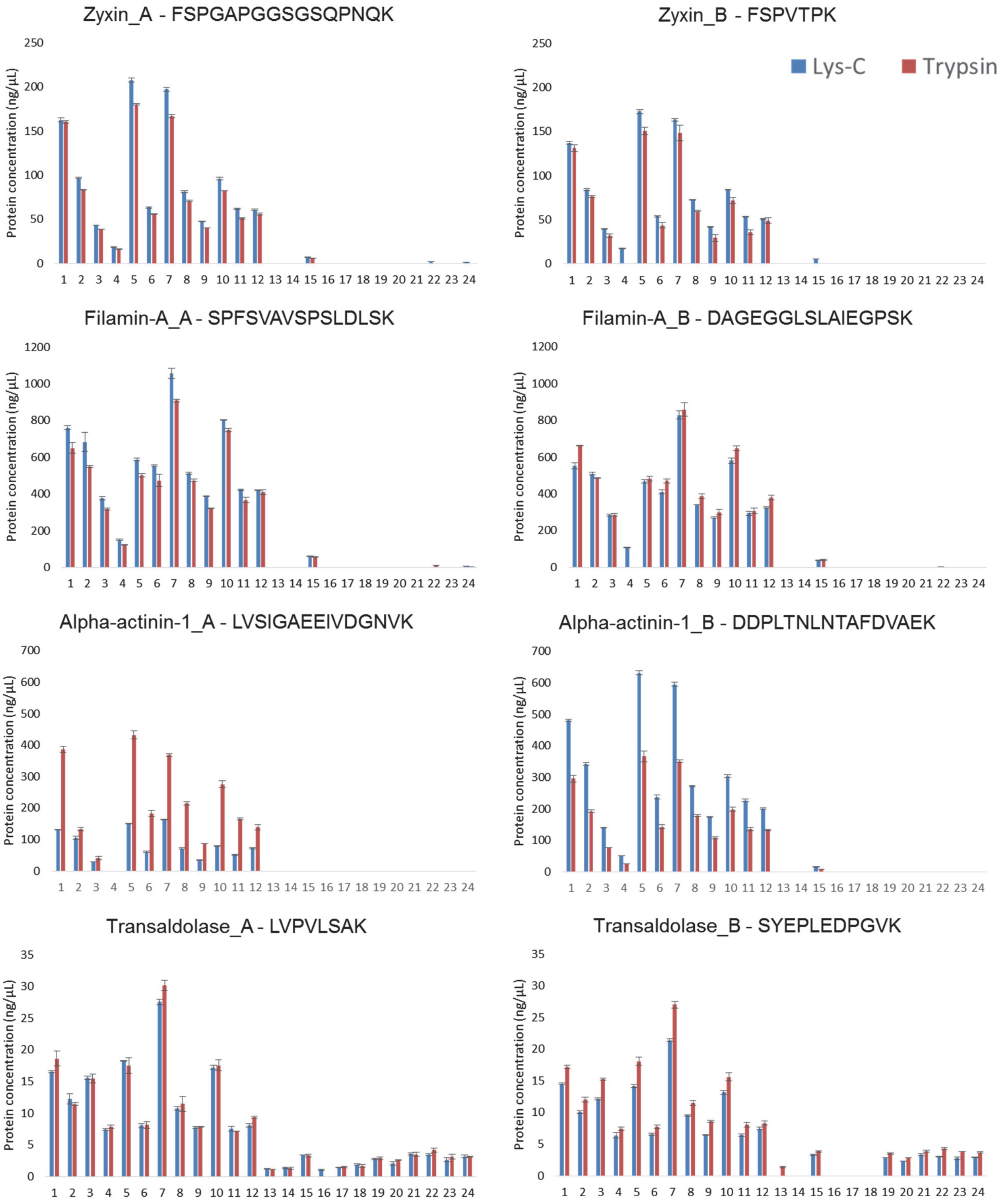


Figure 60: Determination of the accurate concentration of four lung cancer biomarkers: in blue Lys-C and red tryptic plasma digests. Patients diagnosed with NSCLC (n°1 to 12) and healthy people (n°13 to 24).

4. Conclusion

In this study, the accurate quantification of the four lung cancer biomarker candidates was conducted in plasma samples by monitoring eight surrogate peptides (two per protein) generated by two different proteolytic enzymes *i.e.,* Lys-C and trypsin. The accurate quantification was performed using the concatenated polypeptide strategy. The internal standards were designed to target peptides, which are common to both enzymatic digestions. Regardless of the enzymes used, the results confirmed the differential expression of the four targeted proteins in plasma from patients diagnosed with lung cancer compared to healthy volunteers. The quantification of the eight peptides was performed in plasma samples, which were digested in parallel with Lys-C and trypsin. It turned out that a similar quantitative performance was obtained for the Lys-C and the standard tryptic digestion for most of the peptides in terms of digestion efficiency and reproducibility. This experiment also demonstrated that Lys-C proteolysis usually involved in identification studies is well suited also for quantitative assays. The use of Lys-C will allow parts of proteins which are not accessible to trypsin, to be targeted for instance to quantify mutated proteins. This proof-of-principle experiment points out the advantages of the approach. Lys-C is an enzyme with high efficiency, it yields fewer peptides with larger sequences as compared to trypsin thus reducing the background density. Lys-C peptides present good mass spectrometry properties for both ionization and fragmentation (sometimes requires optimization). The concomitant use of isotopically-labeled concatenated polypeptides allows accurate quantification (one point) by leveraging the labeled reporter peptide.

Material & methods

For all the biological samples used during the PhD project the consent of each patient was obtained by filling the informed consent form approved by the Comité National d'Ethique de Recherche.

CHAPTER I

The experimental conditions used in this chapter are described in detail in the published article *Protein quantification using a cleavable reporter peptide* by E. Duriez, S. Trévisiol, B. Domon, J Proteome Res. 2015;14(2):728-37 (Annex 3).

CHAPTER II

The experimental conditions used in this chapter are described in detail in the recently submitted article *Evaluation of alternative enzymes to trypsin for enhancing proteomics analysis* by S. Trévisiol, D. Ayoub, A. Lesur, S. Gallien, B. Domon, recently submitted for publication (Annex 4). The additional experiments are described here.

Chemical modifications of protein lysine residues

Chemicals and reagents

All the chemicals and reagents used in this chapter were purchased from Sigma-Aldrich (Saint Louis, MO, USA) except for sulfosuccinimidyl acetate (sulfo-NHS-acetate) which was purchased from Pierce (Carlsbad, CA, USA). Melittin was provided by Sigma-Aldrich.

Dimethylation

The dimethylation of lysine residues of melittin was performed under agitation (800 rpm) at 37°C during 16h in the presence of 20mM formaldehyde (CH_2O) and 10mM sodium cyanoborohydride ($NaCNBH_3$) in 50 mM HEPES buffer at pH 6-7. After 16h, the excess of reagents was quenched using 100mM ammonium bicarbonate and the proteins are precipitated with acetone (ratio solution volume / acetone volume: 1/10) overnight at -80°C. Subsequently, the supernatant was removed and proteins resuspended in 50 mM HEPES buffer at pH 8-9. Arg-C -like peptides were generated after an overnight tryptic digestion performed under agitation (1000 rpm) at 37°C with an enzyme substrate ratio of 1:20. Before the MS1 analyses on a Velos mass spectrometer samples were desalted using a C_{18} solid phase extraction cartridges (SPE) and dried using a SpeedVac.

Acetylation

Melittin protein was denaturated with 4M guanidine hydrochloride in a 200mM sodium phosphate buffer at pH 8. Cysteine residues were reduced with 3mM tris(2-carboxyethyl)phosphine hydrochloride (TCEP HCl) under agitation (850 rpm) during 1h at 30°C. After that, the sample was diluted with water in order to have a final concentration of 1 M guanidinium hydrochloride and 50 mM sodium phosphate buffer at pH 8. The acetylation of lysine amines was performed with 10mM sulfosuccinimidyl acetate (sulfo-NHS-acetate) in 50 mM sodium phosphate buffer at pH 8 under agitation (850 rpm) during 2h at 30°C. To avoid partial acetylation of other amino acids, such as serines and threonines, the acetylation reaction was reversed using 40 mM hydroxylamine in 50 mM sodium phosphate buffer at pH 8 under agitation (850 rpm) during 10 min at 30°C. The excess reagent was quenched with 20 mM glycine. Before the MS1 analyses on a Velos mass spectrometer, samples were desalted using C18 solid phase extraction cartridges (SPE) and dried using a SpeedVac.

CHAPTER III

Chemicals and reagents

All the chemicals and reagents used in this chapter were purchased from Sigma-Aldrich (Saint Louis, MO, USA) except for RapiGest surfactant which was purchased from Waters (Manchester, UK). Trypsin was provided by Promega (Madison, WI, USA) and Lys-C by Wako (Osaka, Japan). Concatenated polypeptides (>97% peptide purity and >99% isotopic enrichment) and the calibrated isotopologues (>99% peptide purity and >99% isotopic enrichment) were purchased from Thermo Fisher Scientific (Ulm, Germany). The twenty four clinical plasma samples (twelve from patients diagnosed with stage IV NSCLC and twelve from healthy people) were provided by the Integrated Biobank of Luxembourg (IBBL).

Plasma depletion and enzymatic digestions

50 µL of each plasma sample was depleted using multiple affinity removal spin cartridges HAS/IgG (Agilent Technologies) to remove the most abundant proteins of plasma (Albumin and IgGs). Depleted plasma was concentrated using 3kDa cutoff ultrafiltration (Vivaspin 500 3K, Sartorius) and resuspended in 50 mM ammonium bicarbonate with 10% acetonitrile. Each depleted plasma sample was thermally denatured by heating during 10 min at 99°C. At room temperature 0.1% Rapigest in 50mM ammonium bicarbonate, 10% acetonitrile was added to denature the depleted plasma samples. Cysteines were reduced using 10 mM dithiothreitol

(DTT) at 50°C, 50 min followed by alkylation of the thiol groups using 25 mM iodoacetamide (IAA) for 50 min at room temperature in the dark and to finish, the excess of alkylating agents was quenched with 3mM DTT. Plasma protein concentration is roughly 70 mg/mL. The depletion removes around 70% of the total protein amount, thus in 50 µL there is approximately 1050 µg of protein. Each plasma sample was split in two equal parts and digested in parallel in Rapigest 0.1% with trypsin and with Lys-C at pH 8, 12h at 37°C with a ratio E:S 1:100 and 1/40, respectively. A second proteolysis at 37°C was performed for 2h using a ratio E:S 1:100 and 1/263 for Lys-C and trypsin, respectively. After proteolysis the samples were acidified with 10% formic acid to reach pH 2-3 to precipitate Rapigest. Plasma digests were desalted using solid phase extraction cartridges (Sep-Pak C18, Waters, Milford, MA) and dried using a vacuum concentrator. Dried samples were stored at −20 °C before addition of the calibrated internal standards and LC−MS analysis.

Concatenated polypeptide digestion

The eight concatenated polypeptides were digested individually with trypsin overnight in 50 mM ammonium bicarbonate buffer pH 8 at 37°C using a ratio E:S 1/20. After proteolysis the samples were acidified with 10% formic acid to reach pH 2-3. The different digests were desalted using solid phase extraction cartridges (Sep-Pak C18) and dried using a vacuum concentrator. Dried samples were stored at −20 °C before use.

Concatenated polypeptide calibration

Each concatenated polypeptide digest was resuspended with 0.1% formic acid in water. The digests from FSPGAPGGSGSQPNQKAALPAAFK, FSPVTPKAALPAAFK, SPFSVAVSPSLDLSKAALPAAFK and DAGEGGLSLAIEGPSKAALPAAFK were mixed to make the set 1 and LVSIGAEEIVDGNVKAALPAAFK, DDPLTNLNTAFDVAEKAALPAAFK, LVPVLSAKAALPAAFK and SYEPLEDPGVKAALPAAFK were mixed to constitute the set 2. A determined amount of the calibrated isotopologues AALPAAFK (45 amol), AALPAAFK (405 amol), AALPAAFK (3645 amol), and AALPAAFK (32805 amol) was added to each set. The two sets supplemented with internal standards were analyzed by LC-MS on a Q-Exactive Plus mass spectrometer in PRM mode by monitoring the eight isotopologues and the four ISPs.

Liquid chromatography and mass spectrometry

An Ultimate 3000 RSLCnano HPLC system (Thermo Fisher Scientific, San Diego, CA, USA) was used to perform liquid chromatography separations. For each analysis, one µL of peptide mixture was loaded on a trap column (Acclaim PepMap 2 cm × 75 µm i.d., C18, 3 µm, 100 A; Thermo Fisher Scientific) at 5 µL/min with a 1 % acetonitrile + 0.05 % trifluoroacetic acid solution during 3 min. A 0.3 µL/min flow rate was applied and the separations were performed on an analytical column Acclaim PepMap RSLC 15 cm × 75 µm i.d., C18, 2 µm, 100 A (Thermo Fisher Scientific) using a linear gradient of solvent B (acetonitrile + 0.1 % formic acid) into solvent A (water + 0.1 % formic acid). Two different gradients were used. One for the calibration of ISP from 2 % to 35 % B in 33 minutes and a second, for plasma analyses from 2 % to 35 % B in 66 minutes. Both gradients were followed by 4 minutes at 90% B before an equilibration step at 2 % B for 9 minutes.

In the two experiments, the chromatographic system was coupled with a quadrupole-orbitrap mass spectrometer (Q-Exactive Plus, Thermo Fisher Scientific). The acquisition method contains two scan events, a full scan performed in the 300-1500 m/z mass range, with 17500 resolution at 200 m/z, 1e6 AGC target and maximum fill times of 50 ms, and a PRM scan acquired at the 35000 resolution at 200 m/z, 1e6 AGC target, 2 m/z isolation window and a maximum fill time of 100 ms. In the ISP calibration experiment, isotopologue peptides and ISPs were targeted. Fragmentation was performed with a nCE of 25 for the isotopologues, 20 for FSPGAPGGSGSQPNQK, 25 for FSPVTPK, 20 for SPFSVAVSPSLDLSK, 20 for DAGEGGLSLAIEGPSK, 20 for LVSIGAEEIVDGNVK, 20 for DDPLTNLNTAFDVAEK, 28 for LVPVLSAK and 17 for SYEPLEDPGVK. For plasma analyses, the eight pairs of IDPs/endogenous peptides were targeted in ±10 min retention time windows using the nCE for fragmentation previously mentioned. Data extraction to build calibration curves and for the quantitative analyses of lung cancer biomarker candidates were performed using Pinpoint (v1.2 Thermo Fisher Scientific) with a mass tolerance of 10 ppm.

Conclusion & Outlook

Most diseases, in particular cancer, desperately need better markers, either for early detection or for stratifying patients in order to guide treatments. The evaluation of putative biomarkers requires the development of sensitive, robust and accurate analytical methods. During the past decades, bottom-up targeted MS-based proteomics methods have rapidly promoted themselves as cornerstone tools for the quantification of low abundant proteins in complex samples as encountered in biomedical research.

Traditional proteomic approaches for accurate protein quantification rely on the analysis of tryptic peptides, used as surrogates of the proteins of interest, using high purity, calibrated, stable isotope-labeled peptides as internal standards. Although this strategy has already demonstrated its strength for protein identification and quantification in complex biological samples, it is also subjected to some limitations *i.e.,* the incomplete protein sequence coverage and the generation of higher complexity samples due to the use of trypsin. Furthermore, the accuracy of the quantification depends on the quality of the internal standards, typically calibrated only once during manufacturing.

The objective of this work was to develop alternative proteomics approaches to, first, improve the accuracy of the quantitative analyses by using a new type of internal standard, and second, to enhance protein coverage by the generation of fewer peptides, and to reduce sample complexity by employing alternative enzymes to trypsin. Quantitative results are highly dependent on the quality of the used standards. The strategy developed in this study involved the use of concatenated polypeptide standards containing a cleavable reporter peptide which allows the calibration, and subsequence recalibration, of the internal standards prior each analysis. This strategy results in an improvement in accuracy and reliability of the measurement. The amino acid sequence of the cleavable reporter was designed to ensure a high digestion efficiency of the concatenated polypeptides regardless of the isotopically-labeled signature peptides. The calibration is based on the simultaneous measurement of isotopologues in a single LC-MS run, to avoid bias during the calibration process. The constituents of the calibration mixture were selected with sufficient mass differences to avoid overlap between their isotopic distribution patterns.

The bottom-up proteomic approaches, commonly based on a tryptic digestion, generate peptides that are well suited for LC-MS/MS analyses. Nevertheless, due to the uneven distribution of the cleavage sites within the proteome, a large number of non-specific small peptides is generated during tryptic digestion, resulting in a loss of proteome sequence coverage and generation of a more complex background. This study highlighted that alternative enzymes to trypsin were able to access the sequence parts lost during tryptic digestion, which is essential to gain important information such as PTMs, mutations or to distinguish other isoforms. Moreover, enzymes which have different cleavage sites as trypsin, such as Glu-C, have more capabilities to recover the missing segments. It was also demonstrated that enzymes with a single cleavage site, such as Lys-C, generate a lower density background as compared to trypsin, which can be beneficial for the selectivity and consequently the sensitivity of quantitative experiments.

It was also shown that the optimization of collision energies for non-tryptic peptides in order to produce a few numbers of intense fragments can also improve the sensitivity of quantitative experiments, regardless of the type of digestion used.

In the last part of the study the methodologies developed for accurate quantification were applied to measure the concentration of alpha-actinin, zyxin, transaldolase and filamin A in plasma. The concatenated polypeptide approach was used and confirmed the differential expression of these four lung cancer biomarker candidates in plasma from patients diagnosed with lung cancer as compared to plasma from healthy volunteers. Moreover, the quantification of these proteins using Lys-C and trypsin was conducted in parallel, demonstrating that the Lys-C digestion previously employed for protein characterization is also well suited for quantitative studies.

In the field of oncology, the lack of new biomarkers has emerged as one of the main concerns of physicians due to the need of reliable markers for early detection of, and the discrimination between many types of cancers. The evaluation process of the large number of putative markers typically identified in the discovery phase of a biomarker project is divided into three phases. First, the detection of the targets in bodily fluids. Second, the verification to assess sensitivity and specificity. The last step, the validation to evaluate the analytical performances of the optimized assay with a limited number of biomarker candidates to show clinical utility. Over the past decades, extensive lists of putative markers have been established. However, in spite of this wealth of candidates, few have been verified and even less have been validated for the use in the clinic. This has prompted us to rethink the evaluation phase process as described in Figure 61 in order to look more in depth to the protein sequences of targets, such as specific regions containing PTMs, and/or mutations and deletions to discriminate protein variants which could be better disease markers. In this translational workflow for biomarker evaluation, the detection of putative biomarkers derived from discovery studies is conducted in order to determine their differential expression between patients affected by a particular disease and healthy individuals in body fluids such as plasma. Based on the type of isoforms targeted, canonical sequence, splice variants or PTMs, the most appropriate enzyme is determined based on the amino acid sequence. Further, the detectability of peptide targets has to be established in body fluids and is required to show a differential expression. In the verification phase, peptides of interest are precisely quantified using isotopically-labeled signature peptides to assess the sensitivity and specificity of the markers individually or as a panel after a prior optimization of the MS-acquisition parameters and the determination of the reference MS/MS spectra. In the (technical) validation phase, the performances of the assay are evaluated in the optimized analytical conditions (the linearity range, the limit of quantification, precision and accuracy of the measurements) using calibration curves made with isotopically-labeled signature peptides prior calibration with the concatenated polypeptide approach. The limited number of marker candidates which pass this evaluation can be translated into a preliminary clinical assay to assess the selectivity and specificity of the panel on a large cohort of samples. It can be anticipated that in the near future MS-based assays, which can be easily multiplexed, will become generally accepted for clinical assays. Especially for the detection of analytes where antibodies are not available, or are not specific enough to distinguish protein isoforms such PTMs or single point mutations, to develop ELISA tests.

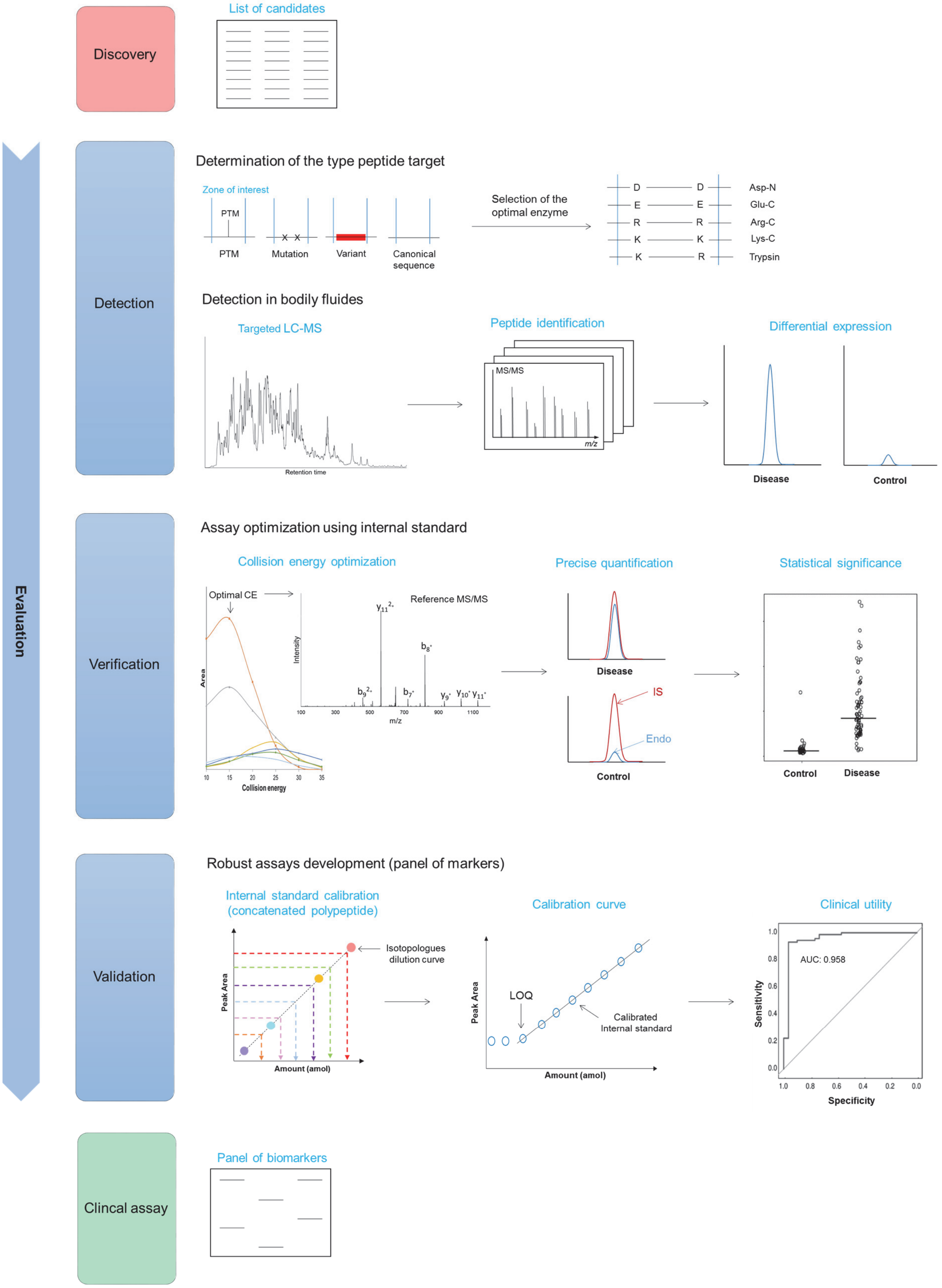


Figure 61: MS-based translational workflow for biomarker evaluation

Bibliography

1. Stewart, B., and Wild, C. P. (2014) World Cancer Report 2014. *The International Agency for Research on Cancer (IARC)*
2. Institute, N. C. (2015) What Is Cancer?
3. Nowell, P. C. (1976) The clonal evolution of tumor cell populations. *Science* 194, 23-28
4. Hanahan, D., and Weinberg, R. A. (2011) Hallmarks of cancer: the next generation. *Cell* 144, 646-674
5. Biesalski, H. K., De Mesquita, B. B., Chesson, A., Chytil, F., Grimble, R., Hermus, R. J. J., Köhrle, J., Lotan, R., Norpoth, K., Pastorino, U., and Thurnham, D. (1998) European Consensus Statement on Lung Cancer: Risk factors and prevention. Lung Cancer Panel. *CA: A Cancer Journal for Clinicians* 48, 167-176
6. Alan Rodgman, T. A. P. (2008) The Chemical Components of Tobacco and Tobacco Smoke. *CRC Press*
7. Catelinois, O., Rogel, A., Laurier, D., Billon, S., Hemon, D., Verger, P., and Tirmarche, M. (2006) Lung Cancer Attributable to Indoor Radon Exposure in France: Impact of the Risk Models and Uncertainty Analysis. *Environmental health perspectives* 114, 1361-1366
8. Hoek, G., and Raaschou-Nielsen, O. (2014) Impact of fine particles in ambient air on lung cancer. *Chinese Journal of Cancer* 33, 197-203
9. A. Stevens, J. L., C. Gompel (1997) *Pathologique générale et spéciale*, De Boeck
10. Subramanian, J., and Govindan, R. (2007) Lung cancer in never smokers: a review. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 25, 561-570
11. ARC, A. p. l. R. s. l. C.-. Brochure les cancers du poumon.
12. Leslie H. Sobin, M. K. G., Christian Wittekind (2009) *TNM Classification of Malignant Tumours, 7th Edition*, Union for International Cancer Control
13. Atkinson, A. J., Colburn, W. A., DeGruttola, V. G., DeMets, D. L., Downing, G. J., Hoth, D. F., Oates, J. A., Peck, C. C., Schooley, R. T., Spilker, B. A., Woodcock, J., and Zeger, S. L. (2001) Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clin Pharmacol Ther* 69, 89-95
14. Bhatt, A. N., Mathur, R., Farooque, A., Verma, A., and Dwarakanath, B. S. (2010) Cancer biomarkers - current perspectives. *The Indian journal of medical research* 132, 129-149
15. Manoj, K. S., K Sarin (2009) Biomarkers of diseases in medicine. *Indian academy of sciences*
16. Brower, V. (2011) Biomarkers: Portents of malignancy. *Nature* 471, S19-S21
17. Duffy, M. J., Duggan, C., Keane, R., Hill, A. D. K., McDermott, E., Crown, J., and O'Higgins, N. (2004) High Preoperative CA 15-3 Concentrations Predict Adverse Outcome in Node-Negative and Node-Positive Breast Cancer: Study of 600 Patients with Histologically Confirmed Breast Cancer. *Clinical chemistry* 50, 559-563
18. Institute, N. C. (2011) Tumor Markers.
19. Neagu, M., Constantin, C., Tanase, C., and Boda, D. (2011) Patented Biomarker Panels in Early Detection of Cancer. *Recent Patents on Biomarkers* 1, 10-24
20. Rifai, N., Gillette, M. A., and Carr, S. A. (2006) Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat Biotech* 24, 971-983

21. Domon, B., and Gallien, S. (2015) Recent advances in targeted proteomics for clinical applications. *PROTEOMICS – Clinical Applications* 9, 423-431

22. Kim, Y. J., Sertamo, K., Pierrard, M.-A., Mesmin, C., Kim, S. Y., Schlesser, M., Berchem, G., and Domon, B. (2015) Verification of the Biomarker Candidates for Non-small-cell Lung Cancer Using a Targeted Proteomics Approach. *Journal of proteome research* 14, 1412-1419

23. Latterich, M., and Schnitzer, J. E. (2011) Streamlining biomarker discovery. *Nat Biotech* 29, 600-602

24. Hanash, S. M., Pitteri, S. J., and Faca, V. M. (2008) Mining the plasma proteome for cancer biomarkers. *Nature* 452, 571-579

25. Nanjappa, V., Thomas, J. K., Marimuthu, A., Muthusamy, B., Radhakrishnan, A., Sharma, R., Ahmad Khan, A., Balakrishnan, L., Sahasrabuddhe, N. A., Kumar, S., Jhaveri, B. N., Sheth, K. V., Kumar Khatana, R., Shaw, P. G., Srikanth, S. M., Mathur, P. P., Shankar, S., Nagaraja, D., Christopher, R., Mathivanan, S., Raju, R., Sirdeshmukh, R., Chatterjee, A., Simpson, R. J., Harsha, H. C., Pandey, A., and Prasad, T. S. K. (2014) Plasma Proteome Database as a resource for proteomics research: 2014 update. *Nucleic acids research* 42, D959-D965

26. Anderson, N. L., and Anderson, N. G. (2002) The human plasma proteome: history, character, and diagnostic prospects. *Molecular & cellular proteomics : MCP* 1, 845-867

27. Schiess, R., Wollscheid, B., and Aebersold, R. (2009) Targeted Proteomic Strategy for Clinical Biomarker Discovery. *Molecular oncology* 3, 33-44

28. Wilkins, M. R., Gasteiger, E., Gooley, A. A., Herbert, B. R., Molloy, M. P., Binz, P.-A., Ou, K., Sanchez, J.-C., Bairoch, A., Williams, K. L., and Hochstrasser, D. F. (1999) High-throughput mass spectrometric discovery of protein post-translational modifications1. *Journal of molecular biology* 289, 645-657

29. Godovac-Zimmermann, J., and Brown, L. R. (2001) Perspectives for mass spectrometry and functional proteomics. *Mass spectrometry reviews* 20, 1-57

30. Chandramouli, K., and Qian, P.-Y. (2009) Proteomics: Challenges, Techniques and Possibilities to Overcome Biological Sample Complexity. *Human Genomics and Proteomics : HGP* 2009, 239204

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

32. James, P. (1997) Protein identification in the post-genome era: the rapid rise of proteomics. *Quarterly reviews of biophysics* 30, 279-331

33. Hoffmann Edmond , S. V. (2007) *Mass Spectrometry: Principles and Applications, 3rd Edition*

34. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* 246, 64-71

35. Karas, M., and Hillenkamp, F. (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical chemistry* 60, 2299-2301

36. Domon, B., and Aebersold, R. (2006) Mass spectrometry and protein analysis. *Science* 312, 212-217

37. Zhang, Y., Fonslow, B. R., Shan, B., Baek, M.-C., and Yates, J. R. (2013) Protein Analysis by Shotgun/Bottom-up Proteomics. *Chemical Reviews* 113, 2343-2394

38. Liumbruno, G., D'Alessandro, A., Grazzini, G., and Zolla, L. (2010) Blood-related proteomics. *Journal of proteomics* 73, 483-507

39. Pieper, R., Su, Q., Gatlin, C. L., Huang, S.-T., Anderson, N. L., and Steiner, S. (2003) Multi-component immunoaffinity subtraction chromatography: An innovative step towards a comprehensive survey of the human plasma proteome. *Proteomics* 3, 422-432

40. Smith, M. P. W., Wood, S. L., Zougman, A., Ho, J. T. C., Peng, J., Jackson, D., Cairns, D. A., Lewington, A. J. P., Selby, P. J., and Banks, R. E. (2011) A systematic analysis of the effects of increasing degrees of serum immunodepletion in terms of depth of coverage and other key aspects in top-down and bottom-up proteomic analyses. *Proteomics* 11, 2222-2235

41. Rodriguez, J., Gupta, N., Smith, R. D., and Pevzner, P. A. (2008) Does Trypsin Cut Before Proline? *Journal of proteome research* 7, 300-305

42. Polgár, L. (2005) The catalytic triad of serine peptidases. *Cell. Mol. Life Sci.* 62, 2161-2172

43. Donald Voet, J. G. V., Charlotte W. Pratt (2008) Principles of Biochemistry 3rd Edition.

44. Guo, X., Trudgian, D. C., Lemoff, A., Yadavalli, S., and Mirzaei, H. (2014) Confetti: A Multiprotease Map of the HeLa Proteome for Comprehensive Proteomics. *Molecular & Cellular Proteomics* 13, 1573-1584

45. Laskay, Ü. A., Lobas, A. A., Srzentić, K., Gorshkov, M. V., and Tsybin, Y. O. (2013) Proteome Digestion Specificity Analysis for Rational Design of Extended Bottom-up and Middle-down Proteomics Experiments. *Journal of proteome research* 12, 5558-5569

46. Wang, Q., Chaerkady, R., Wu, J., Hwang, H. J., Papadopoulos, N., Kopelovich, L., Maitra, A., Matthaei, H., Eshleman, J. R., Hruban, R. H., Kinzler, K. W., Pandey, A., and Vogelstein, B. (2011) Mutant proteins as cancer-specific biomarkers. *Proceedings of the National Academy of Sciences* 108, 2444-2449

47. Tsiatsiani, L., and Heck, A. J. R. (2015) Proteomics beyond trypsin. *FEBS Journal* 282, 2612-2626

48. Norioka, S., Ohta, S., Ohara, T., Lim, S. I., and Sakiyama, F. (1994) Identification of three catalytic triad constituents and Asp-225 essential for function of lysine-specific serine protease, Achromobacter protease I. *Journal of Biological Chemistry* 269, 17025-17029

49. Sørensen, S. B., Sørensen, T. L., and Breddam, K. (1991) Fragmentation of proteins by S. aureus strain V8 protease: Ammonium bicarbonate strongly inhibits the enzyme but does not improve the selectivity for glutamic acid. *FEBS letters* 294, 195-197

50. Auld, D. S. (2013) Chapter 78 - Catalytic Mechanisms for Metallopeptidases. In: Salvesen, N. D. R., ed. *Handbook of Proteolytic Enzymes*, pp. 370-396, Academic Press

51. Ogle, J. D., and Tytell, A. A. (1953) The activity of Clostridium histolyticum proteinase on synthetic substrates. *Archives of Biochemistry and Biophysics* 42, 327-336

52. E. Labrou, N. (2013) Chapter 521 - Clostripain. In: Salvesen, N. D. R., ed. *Handbook of Proteolytic Enzymes*, pp. 2323-2327, Academic Press

53. Olsen, J. V., Ong, S. E., and Mann, M. (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Molecular & cellular proteomics : MCP* 3, 608-614

54. Holmes, J. L. (1984) F. W. McLafferty (ED.) Tandem msss spectrometry. Wiley–Interscience, New York, 1983. Price US $46.20. *Organic Mass Spectrometry* 19, 530-530

55. Johnson, J. V., Yost, R. A., Kelley, P. E., and Bradford, D. C. (1990) Tandem-in-space and tandem-in-time mass spectrometry: triple quadrupoles and quadrupole ion traps. *Analytical chemistry* 62, 2162-2172

56. Vestal, M. L., and Campbell, J. M. (2005) Tandem Time‐of‐Flight Mass Spectrometry. *Methods in enzymology*, pp. 79-108, Academic Press

57. Peterson, A. C., Russell, J. D., Bailey, D. J., Westphall, M. S., and Coon, J. J. (2012) Parallel Reaction Monitoring for High Resolution and High Mass Accuracy Quantitative, Targeted Proteomics. *Molecular & Cellular Proteomics* 11, 1475-1488

58. Gallien, S., Duriez, E., Crone, C., Kellmann, M., Moehring, T., and Domon, B. (2012) Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer. *Molecular & cellular proteomics : MCP* 11, 1709-1723

59. Shevchenko, A., Loboda, A., Shevchenko, A., Ens, W., and Standing, K. G. (2000) MALDI Quadrupole Time-of-Flight Mass Spectrometry: A Powerful Tool for Proteomic Research. *Analytical chemistry* 72, 2132-2141

60. Douglas, D. J., Frank, A. J., and Mao, D. (2005) Linear ion traps in mass spectrometry. *Mass spectrometry reviews* 24, 1-29

61. Roepstorff, P., and Fohlman, J. (1984) Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical mass spectrometry* 11, 601

62. Biemann, K. (1988) Contributions of mass spectrometry to peptide and protein structure. *Biomedical & environmental mass spectrometry* 16, 99-111

63. Johnson, R. S., Martin, S. A., and Biemann, K. (1988) Collision-induced fragmentation of (M + H)+ ions of peptides. Side chain specific sequence ions. *International Journal of Mass Spectrometry and Ion Processes* 86, 137-154

64. Mitchell Wells, J., and McLuckey, S. A. (2005) Collision‐Induced Dissociation (CID) of Peptides and Proteins. In: Burlingame, A. L., ed. *Methods in enzymology*, pp. 148-185, Academic Press

65. Papayannopoulos, I. A. (1995) The interpretation of collision-induced dissociation tandem mass spectra of peptides. *Mass spectrometry reviews* 14, 49-73

66. Dongré, A. R., Jones, J. L., Somogyi, Á., and Wysocki, V. H. (1996) Influence of Peptide Composition, Gas-Phase Basicity, and Chemical Modification on Fragmentation Efficiency: Evidence for the Mobile Proton Model. *Journal of the American Chemical Society* 118, 8365-8374

67. Paizs, B., and Suhai, S. (2005) Fragmentation pathways of protonated peptides. *Mass spectrometry reviews* 24, 508-548

68. Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J., and Hunt, D. F. (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America* 101, 9528-9533

69. Domon, B., and Aebersold, R. (2010) Options and considerations when selecting a quantitative proteomics strategy. *Nat Biotech* 28, 710-721

70. Yates, J. R., 3rd (1998) Mass spectrometry and the age of the proteome. *Journal of mass spectrometry : JMS* 33, 1-19

71. Hebert, A. S., Richards, A. L., Bailey, D. J., Ulbrich, A., Coughlin, E. E., Westphall, M. S., and Coon, J. J. (2014) The One Hour Yeast Proteome. *Molecular & cellular proteomics : MCP* 13, 339-347

72. Gallien, S., Duriez, E., and Domon, B. (2011) Selected reaction monitoring applied to proteomics. *Journal of mass spectrometry : JMS* 46, 298-312

73. Hoke Ii, S. H., Morand, K. L., Greis, K. D., Baker, T. R., Harbol, K. L., and Dobson, R. L. M. (2001) Transformations in pharmaceutical research and development, driven by innovations in multidimensional mass spectrometry-based technologies. *International Journal of Mass Spectrometry* 212, 135-196

74. Kostiainen, R., Kotiaho, T., Kuuranne, T., and Auriola, S. (2003) Liquid chromatography/atmospheric pressure ionization–mass spectrometry in drug metabolism studies. *Journal of Mass Spectrometry* 38, 357-372

75. Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S., and Mann, M. (2007) Higher-energy C-trap dissociation for peptide modification analysis. *Nat Meth* 4, 709-712

76. Hawkridge, A. M. (2014) CHAPTER 1 Practical Considerations and Current Limitations in Quantitative Mass Spectrometry-based Proteomics. *Quantitative Proteomics*, pp. 1-25, The Royal Society of Chemistry

77. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotech* 17, 994-999

78. Ross, P. L., Huang, Y. N., Marchese, J. N., 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., and Pappin, D. J. (2004) Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents. *Molecular & Cellular Proteomics* 3, 1154-1169

79. Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., and Hamon, C. (2003) Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. *Analytical chemistry* 75, 1895-1904

80. Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002) Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Molecular & Cellular Proteomics* 1, 376-386

81. Yao, X., Afonso, C., and Fenselau, C. (2003) Dissection of Proteolytic 18O Labeling: Endoprotease-Catalyzed 16O-to-18O Exchange of Truncated Peptide Substrates. *Journal of proteome research* 2, 147-152

82. Miyagi, M., and Rao, K. C. S. (2007) Proteolytic 18O-labeling strategies for quantitative proteomics. *Mass spectrometry reviews* 26, 121-136

83. Johnson, K. L., and Muddiman, D. C. (2004) A method for calculating 16o/18o peptide ion ratios for the relative quantification of proteomes. *Journal of the American Society for Mass Spectrometry* 15, 437-445

84. Bondarenko, P. V., Chelius, D., and Shaler, T. A. (2002) Identification and Relative Quantitation of Protein Mixtures by Enzymatic Digestion Followed by Capillary Reversed-Phase Liquid Chromatography−Tandem Mass Spectrometry. *Analytical chemistry* 74, 4741-4749

85. Liu, H., Sadygov, R. G., and Yates, J. R. (2004) A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics. *Analytical chemistry* 76, 4193-4201

86. Chelius, D., and Bondarenko, P. V. (2002) Quantitative Profiling of Proteins in Complex Mixtures Using Liquid Chromatography and Mass Spectrometry. *Journal of proteome research* 1, 317-323

87. Washburn, M. P., Ulaszek, R. R., and Yates, J. R. (2003) Reproducibility of Quantitative Proteomic Analyses of Complex Biological Mixtures by Multidimensional Protein Identification Technology. *Analytical chemistry* 75, 5054-5061

88. Ludwig, C., and Aebersold, R. (2014) CHAPTER 4 Getting Absolute: Determining Absolute Protein Quantities via Selected Reaction Monitoring Mass Spectrometry. *Quantitative Proteomics*, pp. 80-109, The Royal Society of Chemistry

89. Brun, V., Masselon, C., Garin, J., and Dupuis, A. (2009) Isotope dilution strategies for absolute quantitative proteomics. *Journal of proteomics* 72, 740-749

90. Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W., and Gygi, S. P. (2003) Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proceedings of the National Academy of Sciences* 100, 6940-6945

91. Pratt, J. M., Simpson, D. M., Doherty, M. K., Rivers, J., Gaskell, S. J., and Beynon, R. J. (2006) Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes. *Nat. Protocols* 1, 1029-1043

92. Brun, V., Dupuis, A., Adrait, A., Marcellin, M., Thomas, D., Court, M., Vandenesch, F., and Garin, J. (2007) Isotope-labeled Protein Standards: Toward Absolute Quantitative Proteomics. *Molecular & Cellular Proteomics* 6, 2139-2149

93. Carr, S. A., Abbatiello, S. E., Ackermann, B. L., Borchers, C., Domon, B., Deutsch, E. W., Grant, R. P., Hoofnagle, A. N., Hüttenhain, R., Koomen, J. M., Liebler, D. C., Liu, T., MacLean, B., Mani, D., Mansfield, E., Neubert, H., Paulovich, A. G., Reiter, L., Vitek, O., Aebersold, R., Anderson, L., Bethem, R., Blonder, J., Boja, E., Botelho, J., Boyne, M., Bradshaw, R. A., Burlingame, A. L., Chan, D., Keshishian, H., Kuhn, E., Kinsinger, C., Lee, J. S. H., Lee, S.-W., Moritz, R., Oses-Prieto, J., Rifai, N., Ritchie, J., Rodriguez, H., Srinivas, P. R., Townsend, R. R., Van Eyk, J., Whiteley, G., Wiita, A., and Weintraub, S. (2014) Targeted Peptide Measurements in Biology and Medicine: Best Practices for Mass Spectrometry-based Assay Development Using a Fit-for-Purpose Approach. *Molecular & Cellular Proteomics* 13, 907-917

94. Addona, T. A., Abbatiello, S. E., Schilling, B., Skates, S. J., Mani, D. R., Bunk, D. M., Spiegelman, C. H., Zimmerman, L. J., Ham, A.-J. L., Keshishian, H., Hall, S. C., Allen, S., Blackman, R. K., Borchers, C. H., Buck, C., Cardasis, H. L., Cusack, M. P., Dodder, N. G., Gibson, B. W., Held, J. M., Hiltke, T., Jackson, A., Johansen, E. B., Kinsinger, C. R., Li, J., Mesri, M., Neubert, T. A., Niles, R. K., Pulsipher, T. C., Ransohoff, D., Rodriguez, H., Rudnick, P. A., Smith, D., Tabb, D. L., Tegeler, T. J., Variyath, A. M., Vega-Montoto, L. J., Wahlander, A., Waldemarson, S., Wang, M., Whiteaker, J. R., Zhao, L., Anderson, N. L., Fisher, S. J., Liebler, D. C., Paulovich, A. G., Regnier, F. E., Tempst, P., and Carr, S. A. (2009) Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat Biotech* 27, 633-641

95. Keshishian, H., Addona, T., Burgess, M., Mani, D. R., Shi, X., Kuhn, E., Sabatine, M. S., Gerszten, R. E., and Carr, S. A. (2009) Quantification of Cardiovascular Biomarkers in Patient Plasma by

Targeted Mass Spectrometry and Stable Isotope Dilution. *Molecular & Cellular Proteomics* 8, 2339-2349

96. Kuzyk, M. A., Smith, D., Yang, J., Cross, T. J., Jackson, A. M., Hardie, D. B., Anderson, N. L., and Borchers, C. H. (2009) Multiple Reaction Monitoring-based, Multiplexed, Absolute Quantitation of 45 Proteins in Human Plasma. *Molecular & Cellular Proteomics* 8, 1860-1877

97. Lopez, M. F., Kuppusamy, R., Sarracino, D. A., Prakash, A., Athanas, M., Krastins, B., Rezai, T., Sutton, J. N., Peterman, S., and Nicolaides, K. (2011) Mass Spectrometric Discovery and Selective Reaction Monitoring (SRM) of Putative Protein Biomarker Candidates in First Trimester Trisomy 21 Maternal Serum. *Journal of proteome research* 10, 133-142

98. Addona, T. A., Shi, X., Keshishian, H., Mani, D. R., Burgess, M., Gillette, M. A., Clauser, K. R., Shen, D., Lewis, G. D., Farrell, L. A., Fifer, M. A., Sabatine, M. S., Gerszten, R. E., and Carr, S. A. (2011) A pipeline that integrates the discovery and verification of plasma protein biomarkers reveals candidate markers for cardiovascular disease. *Nat Biotech* 29, 635-643

99. Picotti, P., and Aebersold, R. (2012) Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat Meth* 9, 555-566

100. Holzmann, J., Pichler, P., Madalinski, M., Kurzbauer, R., and Mechtler, K. (2009) Stoichiometry Determination of the MP1−p14 Complex Using a Novel and Cost-Efficient Method To Produce an Equimolar Mixture of Standard Peptides. *Analytical chemistry* 81, 10254-10261

101. Mirzaei, H., McBee, J. K., Watts, J., and Aebersold, R. (2008) Comparative Evaluation of Current Peptide Production Platforms Used in Absolute Quantification in Proteomics. *Molecular & Cellular Proteomics* 7, 813-823

102. Horinek, D., Serr, A., Geisler, M., Pirzer, T., Slotta, U., Lud, S. Q., Garrido, J. A., Scheibel, T., Hugel, T., and Netz, R. R. (2008) Peptide adsorption on a hydrophobic surface results from an interplay of solvation, surface, and intrapeptide forces. *Proceedings of the National Academy of Sciences of the United States of America* 105, 2842-2847

103. Rutherfurd, S. M., and Gilani, G. S. (2001) Amino Acid Analysis. *Current Protocols in Protein Science*, John Wiley & Sons, Inc.

104. Winter, D., Hung, C.-W., Jaskolla, T. W., Karas, M., and Lehmann, W. D. (2012) Enzyme-cleavable tandem peptides for quantitative studies in MS-based proteomics. *Proteomics* 12, 3470-3474

105. Singh, S., Springer, M., Steen, J., Kirschner, M. W., and Steen, H. (2009) FLEXIQuant: A Novel Tool for the Absolute Quantification of Proteins, and the Simultaneous Identification and Quantification of Potentially Modified Peptides. *Journal of proteome research* 8, 2201-2210

106. Yeoun-Jin Kim, E. D., Sébastien Gallien, Bruno Domon (2011) Novel Proteomics Approach for Absolute Quantification Using Polypeptides Containing a Reporter as Internal Standards. *Poster presentation ASMS*

107. Yeoun-Jin Kim, E. D., Sébastien Gallien, Guy Berchem, Bruno Domon (2011) Biomarker detection and quantification in bodily fluids using concatenated reference peptides including a universal reporter. *Poster presentation HUPO*

108. Krokhin, O. V., and Spicer, V. (2010) Predicting Peptide Retention Times for Proteomics. *Current Protocols in Bioinformatics*, John Wiley & Sons, Inc.

109. Krokhin, O. V. (2006) Sequence-Specific Retention Calculator. Algorithm for Peptide Retention Prediction in Ion-Pair RP-HPLC: Application to 300- and 100-Å Pore Size C18 Sorbents. *Analytical chemistry* 78, 7785-7795

110. Thiede, B., Lamer, S., Mattow, J., Siejak, F., Dimmler, C., Rudel, T., and Jungblut, P. R. (2000) Analysis of missed cleavage sites, tryptophan oxidation and N-terminal pyroglutamylation after in-gel tryptic digestion. *Rapid Communications in Mass Spectrometry* 14, 496-502

111. Turowski, M., Yamakawa, N., Meller, J., Kimata, K., Ikegami, T., Hosoya, K., Tanaka, N., and Thornton, E. R. (2003) Deuterium Isotope Effects on Hydrophobic Interactions: The Importance of Dispersion Interactions in the Hydrophobic Phase. *Journal of the American Chemical Society* 125, 13836-13849

112. Julka, S., and Regnier, F. (2004) Quantification in Proteomics through Stable Isotope Coding: A Review. *Journal of proteome research* 3, 350-363

113. Hansen, K. C., Schmitt-Ulms, G., Chalkley, R. J., Hirsch, J., Baldwin, M. A., and Burlingame, A. L. (2003) Mass Spectrometric Analysis of Protein Mixtures at Low Levels Using Cleavable 13C-Isotope-coded Affinity Tag and Multidimensional Chromatography. *Molecular & Cellular Proteomics* 2, 299-314

114. Paret, C., Schön, Z., Szponar, A., and Kovacs, G. (2010) Inflammatory Protein Serum Amyloid A1 Marks a Subset of Conventional Renal Cell Carcinomas with Fatal Outcome. *European urology* 57, 859-866

115. Mattarollo, S. R., and Smyth, M. J. (2010) A novel axis of innate immunity in cancer. *Nat Immunol* 11, 981-982

116. Indovina, P., Marcelli, E., Maranta, P., and Tarro, G. (2011) Lung Cancer Proteomics: Recent Advances in Biomarker Discovery. *International journal of proteomics* 2011, 726869

117. Sung, H.-J., Ahn, J.-M., Yoon, Y.-H., Rhim, T.-Y., Park, C.-S., Park, J.-Y., Lee, S.-Y., Kim, J.-W., and Cho, J.-Y. (2011) Identification and Validation of SAA as a Potential Lung Cancer Biomarker and its Involvement in Metastatic Pathogenesis of Lung Cancer. *Journal of proteome research* 10, 1383-1395

118. Lesur, A., Ancheva, L., Kim, Y. J., Berchem, G., van Oostrum, J., and Domon, B. (2015) Screening protein isoforms predictive for cancer using immunoaffinity capture and fast LC-MS in PRM mode. *PROTEOMICS – Clinical Applications*, n/a-n/a

119. Cedano, J., Aloy, P., Pérez-Pons, J. A., and Querol, E. (1997) Relation between amino acid composition and cellular location of proteins1. *Journal of molecular biology* 266, 594-600

120. Gallien, S., Duriez, E., Demeure, K., and Domon, B. (2013) Selectivity of LC-MS/MS analysis: implication for proteomics experiments. *Journal of proteomics* 81, 148-158

121. Swaney, D. L., Wenger, C. D., and Coon, J. J. (2010) The value of using multiple proteases for large-scale mass spectrometry-based proteomics. *Journal of proteome research* 9, 1323-1329

122. Rayment, I. (1997) [12] Reductive alkylation of lysine residues to alter crystallization properties of proteins. In: Charles W. Carter, Jr., ed. *Methods in enzymology*, pp. 171-179, Academic Press

123. Gidley, M. J., and Sanders, J. K. (1982) Reductive methylation of proteins with sodium cyanoborohydride. Identification, suppression and possible uses of N-cyanomethyl by-products. *Biochemical Journal* 203, 331-334

124. Dottavio-Martin, D., and Ravel, J. M. (1978) Radiolabeling of proteins by reductive alkylation with [14C]formaldehyde and sodium cyanoborohydride. *Analytical biochemistry* 87, 562-565

125. Jentoft, N., and Dearborn, D. G. (1979) Labeling of proteins by reductive methylation using sodium cyanoborohydride. *Journal of Biological Chemistry* 254, 4359-4365

126. Abello, N., Kerstjens, H. A. M., Postma, D. S., and Bischoff, R. (2007) Selective Acylation of Primary Amines in Peptides and Proteins. *Journal of proteome research* 6, 4770-4776

127. Van Damme, P., Van Damme, J., Demol, H., Staes, A., Vandekerckhove, J., and Gevaert, K. (2009) A review of COFRADIC techniques targeting protein N-terminal acetylation. *BMC Proceedings* 3, S6

128. Gevaert, K., Goethals, M., Martens, L., Van Damme, J., Staes, A., Thomas, G. R., and Vandekerckhove, J. (2003) Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat Biotech* 21, 566-569

129. Kuster, B., Schirle, M., Mallick, P., and Aebersold, R. (2005) Scoring proteomes with proteotypic peptide probes. *Nature reviews. Molecular cell biology* 6, 577-583

130. Vaisar, T., and Urban, J. (1996) Probing Proline Effect in CID of Protonated Peptides. *Journal of Mass Spectrometry* 31, 1185-1187

131. Gu, C., Tsaprailis, G., Breci, L., and Wysocki, V. H. (2000) Selective Gas-Phase Cleavage at the Peptide Bond C-Terminal to Aspartic Acid in Fixed-Charge Derivatives of Asp-Containing Peptides. *Analytical chemistry* 72, 5804-5813

132. Kelstrup, C. D., Jersie-Christensen, R. R., Batth, T. S., Arrey, T. N., Kuehn, A., Kellmann, M., and Olsen, J. V. (2014) Rapid and Deep Proteomes by Faster Sequencing on a Benchtop Quadrupole Ultra-High-Field Orbitrap Mass Spectrometer. *Journal of proteome research* 13, 6187-6195

133. Scheltema, R. A., Hauschild, J.-P., Lange, O., Hornburg, D., Denisov, E., Damoc, E., Kuehn, A., Makarov, A., and Mann, M. (2014) The Q Exactive HF, a Benchtop Mass Spectrometer with a Pre-filter, High Performance Quadrupole and an Ultra-High Field Orbitrap Analyzer. *Molecular & Cellular Proteomics*

134. Michalski, A., Damoc, E., Hauschild, J.-P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M., and Horning, S. (2011) Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer. *Molecular & Cellular Proteomics* 10

135. Wan, K., Vidavsky, I., and Gross, M. (2002) Comparing similar spectra: From similarity index to spectral contrast angle. *Journal of the American Society for Mass Spectrometry* 13, 85-88

136. Gallien, S., and Domon, B. (2015) Detection and quantification of proteins in clinical samples using high resolution mass spectrometry. *Methods* 81, 15-23

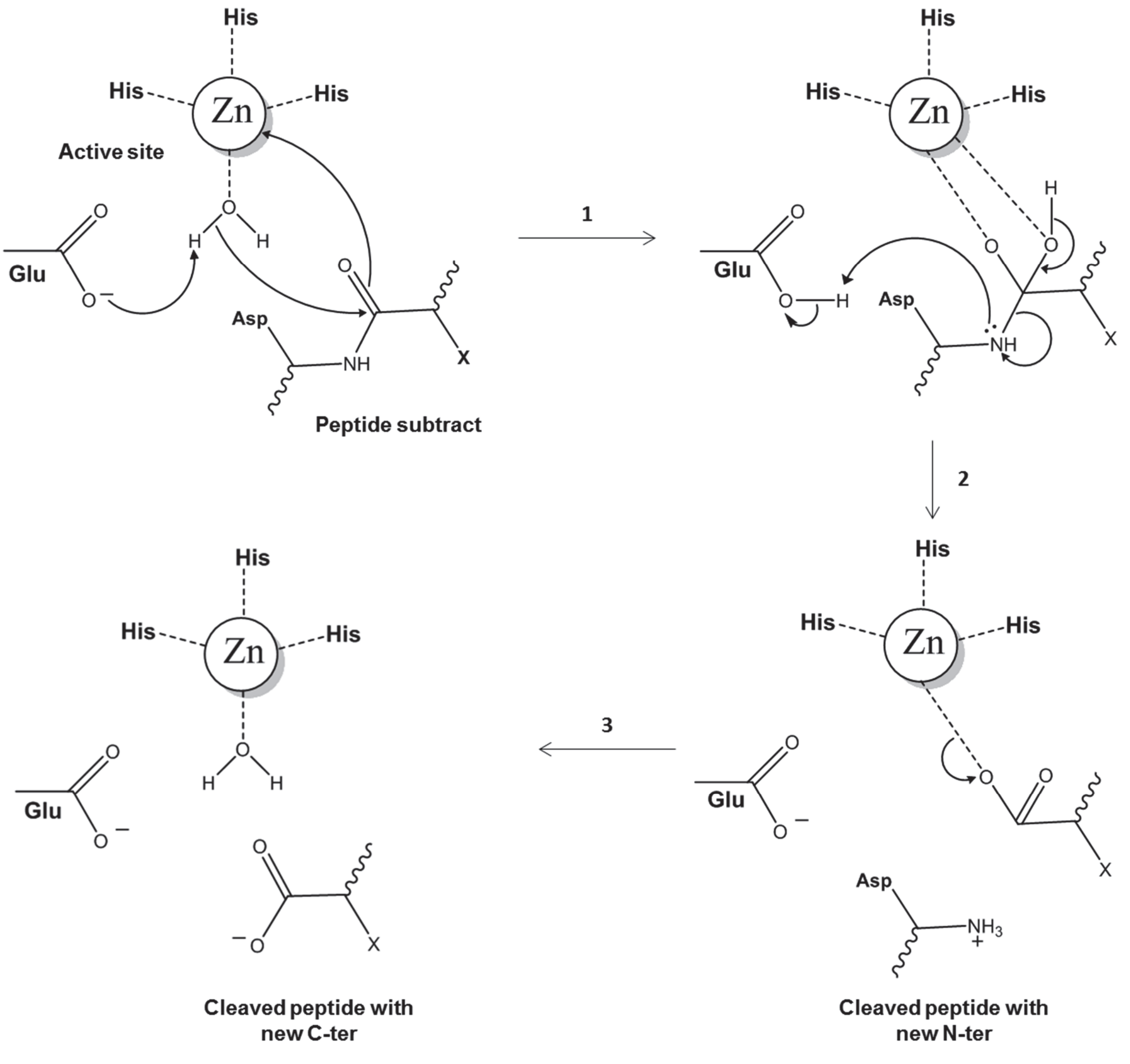
137. Gallien, S., Kim, S. Y., and Domon, B. (2015) Large-Scale Targeted Proteomics Using Internal Standard Triggered-Parallel Reaction Monitoring (IS-PRM). *Molecular & Cellular Proteomics* 14, 1630-1644

138. Gershon, P. D. (2014) Cleaved and Missed Sites for Trypsin, Lys-C, and Lys-N Can Be Predicted with High Confidence on the Basis of Sequence Context. *Journal of proteome research* 13, 702-709

ANNEXES

Annex 1

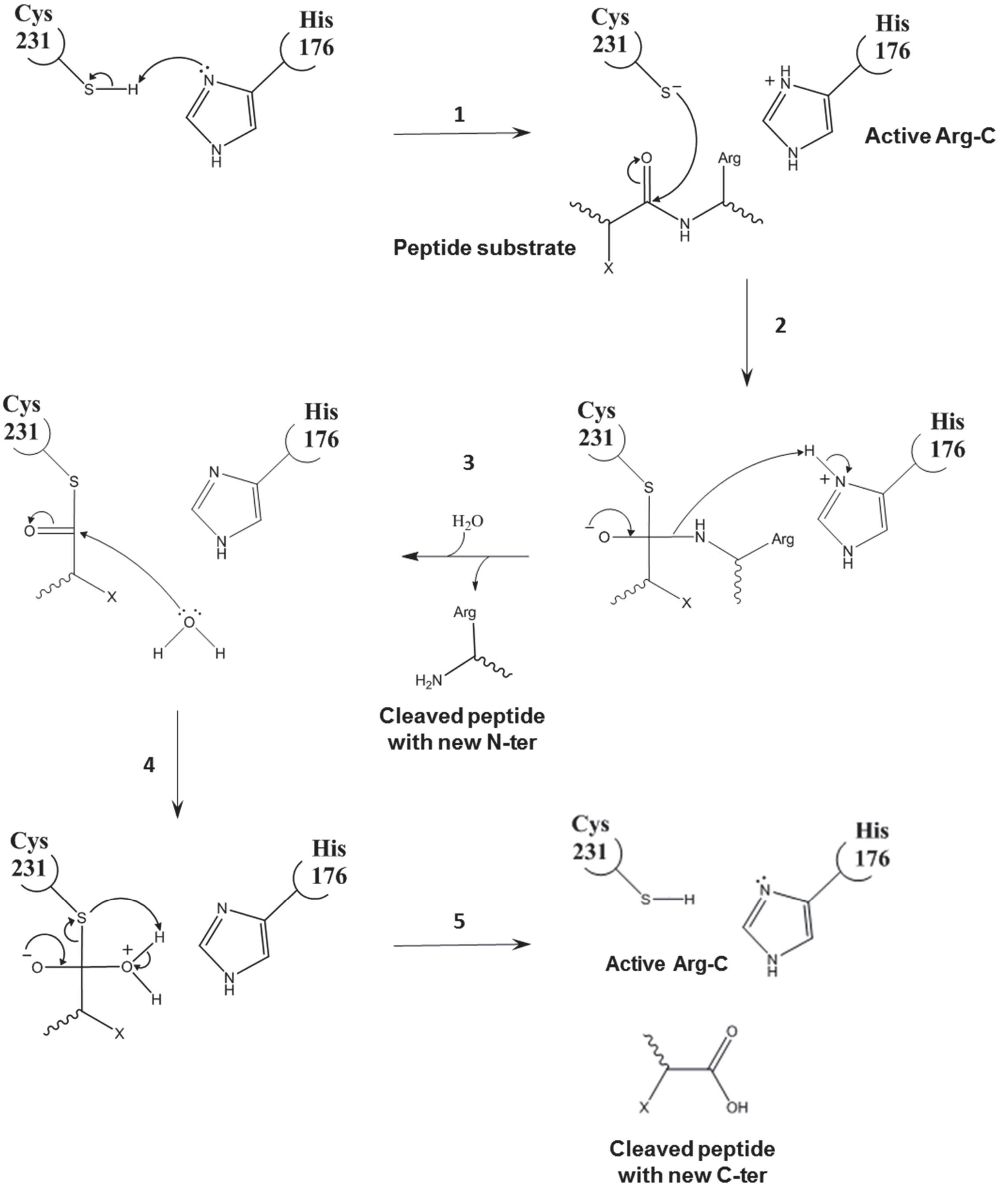
Zinc metalloendopeptidase catalytic mechanism applied to Asp-N (adapted from Auld, Handbook of Proteolytic Enzymes, 3rd Ed)



His
His
His
Zn
Active site
O
O
H
H
Glu
O−
O
Asp
NH
X
Peptide subtract
1
His
His
His
Zn
H
O
O
O
Glu
O
H
Asp
NH
X
2
His
His
His
Zn
Glu
O
O−
O
O
X
Asp
NH_3^+
Cleaved peptide with new N-ter
3
His
His
His
Zn
O
H
H
Glu
O
O−
O
−O
X
Cleaved peptide with new C-ter

Annex 2

Cysteine activated protease catalytic mechanism applied to Arg-C (adapted from Auld, Handbook of Proteolytic Enzymes, 3rd Ed)


Cys 231
His 176
1
2
3
4
5
Active Arg-C
Peptide substrate
Arg
X
H_2O
Cleaved peptide with new N-ter
H_2N
Active Arg-C
Cleaved peptide with new C-ter
OH

Annex 3

Protein quantification using a cleavable reporter peptide

E. Duriez, S. Trévisiol., B. Domon

Journal of Proteome Research, 2015; 2: 728-737

Journal of proteome research

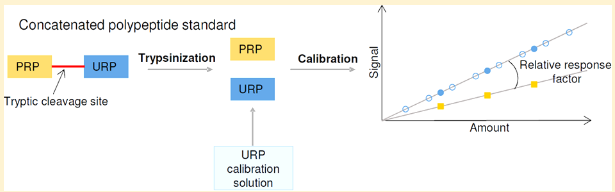
Protein Quantification Using a Cleavable Reporter Peptide

Elodie Duriez, Stephane Trevisiol, and Bruno Domon*

Luxembourg Clinical Proteomics Center, CRP-Santé, 1 A-B rue Thomas Edison, Strassen 1445, Luxembourg

S *Supporting Information*

ABSTRACT: Peptide and protein quantification based on isotope dilution and mass spectrometry analysis are widely employed for the measurement of biomarkers and in system biology applications. The accuracy and reliability of such quantitative assays depend on the quality of the stable-isotope labeled standards. Although the quantification using stable-isotope labeled peptides is precise, the accuracy of the results can be severely biased by the purity of the internal standards, their stability and formulation, and the determination of their concentration. Here we describe a rapid and cost-efficient method to recalibrate stable isotope labeled peptides in a single LC−MS analysis. The method is based on the equimolar release of a protein reference peptide (used as surrogate for the protein of interest) and a universal reporter peptide during the trypsinization of a concatenated polypeptide standard. The quality and accuracy of data generated with such concatenated polypeptide standards are highlighted by the quantification of two clinically important proteins in urine samples and compared with results obtained with conventional stable isotope labeled reference peptides. Furthermore, the application of the UCRP standards in complex samples is described.



KEYWORDS: *mass spectrometry, targeted proteomics, peptide/protein quantification, stable isotope labeled peptides, calibration, standard recalibration, cleavable reporter peptide*

INTRODUCTION

During the past decade, mass-spectrometry-based proteomics has become an essential tool in biological and clinical investigation.[1,2] Most protein quantification approaches of biological samples generally involve trypsin digestion of the endogenous proteins followed by a targeted LC−MS-based analysis of signature peptides indicative of the proteins of interest.[1] An isotope dilution strategy is frequently employed to achieve precise quantification of the proteins of interest using stable isotope-labeled (SIL) standards.[1] Several SIL approaches have been proposed, including: synthetic peptides (SIL peptides), extended-peptides, concatemers of peptides (QconCAT), or full-length proteins (PSAQ).[3−7] Recently, a novel type of SIL peptides (differing by the number of neutrons incorporated and using high-resolution accurate mass spectrometry) has been proposed.[8] In isotope dilution experiments, the accuracy and reliability of the quantitative assay are directly dependent on the quality criteria of the SIL standards, which include isotope incorporation, chemical purity (isotopic purity), and the actual concentration of the SIL standards.[9]

SIL peptides (sometimes referred to as AQUA peptides) are frequently used due to their commercial availability and straightforward applicability.[10−14] However, it is also known that quantification using SIL peptides can be biased, for example, due to a partial uncontrolled loss of the SIL peptides before addition to the biological samples.[15] Several factors can be attributed to a reduction of the initial amounts of SIL peptides. First, the solubilization of the lyophilized SIL peptides may be incomplete.[16,17] Second, peptides in solution are prone to nonspecific adsorption to hydrophobic surfaces or can be subject to a time dependent aggregation. These effects are difficult to predict based solely on the peptide sequence, and thus the formulation buffer, the vials, and storage conditions are critical assay components.[18] Thus, the quantification based on SIL peptides often results in precise but inaccurate results, unless the various pitfalls associated with the method are taken into consideration.

The LC−MS-based quantification of proteins after digestion relies on the initial amount of the SIL peptides indicated by the manufacturer, which is determined either by photometric methods or by quantitative amino acid analysis (AAA).[19] In principle, the SIL peptide concentration should be recalibrated, before each use, to achieve accurate quantitative results. As AAA assays are tedious, time-consuming, costly, and usually performed by an external laboratory requiring significant amount of material (typically >50 μg) for a single analysis, the method is not well-suited for routine recalibration of the SIL peptide concentration. There is a need for methods allowing the recalibration of SIL peptides amount in a routine analysis suitable fashion.

Different methods have been recently proposed for the quantification of SIL standards based on equimolar products generation by proteolysis.[16,20] Synthetic peptides isotopically labeled at the N-terminus and concatenated to conventional SIL peptides by an enzymatic cleavable site have been proposed to determine the stoichiometry of protein complexes.[16] Similarly, the generation of an equimolar mixture upon

Received: July 22, 2014
Published: November 20, 2014

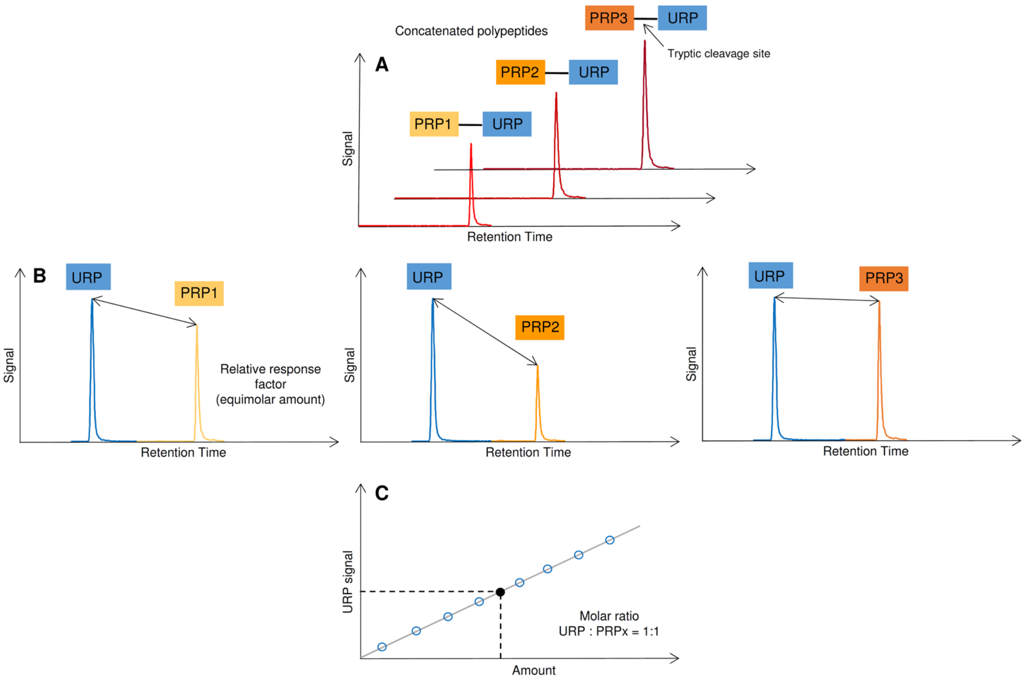


Figure 1. Representation of the concatenated polypeptide standards and the methodology associated. (A) Individual analyses of the polypeptide standards, incorporating a tryptic cleavage site and containing a protein reference peptide (PRP) ligated to a universal reporter peptide (URP). Because the URP sequence is the same for all UCRP standards, multiplexing PRP calibration is possible using different SIL amino acids of the URP. (B) Analyses of the digested standard samples and determination of the relative response factor between the 1:1 stoichiometric released products, URP and PRP. (C) Determination of the URP and PRP amounts using a reverse URP standard calibration curve.

digestion of tandem peptides was used to determine relative MS ionization efficiencies.[20] Full-length protein standards containing an N-terminal enzymatic cleavable peptide have been proposed where the N-terminal peptide, as a surrogate for the whole protein, is quantified.[21] The use of concatenated synthetic polypeptides with a trypsin cleavage site can be further exploited to systematically determine the concentration of the SIL peptide standard of interest upon tryptic digestion. (See Figure 1.) The universal cleavable reporter peptide (UCRP) standard discussed here has the peptide to be quantified fused to a universal reporter peptide (URP), whose concentration can be readily determined and used to calculate the relative response of the protein reference peptide (PRP). Conversely, if the response factors have already been determined, the concentration of the URP allows us to determine the concentration of the PRP of interest. As the URP sequence is identical for all UCRP standards, this peptide can be systematically used to recalibrate any PRP (Figure 1). The methodology described in this study includes the determination of the PRP concentration, using a calibration curve composed of URP isotopologues accurately quantified, and the comparison of the cleavable reporter peptide standards to the results obtained using direct quantification with conventional SIL peptides. The calibration using UCRP standards can be performed either individually or concomitantly if multiple labeling schemes are used.

■ MATERIAL AND METHODS

Synthetic Peptides

Isotopically labeled (SIL) peptides, URP isotopologues, and UCRP standards were synthesized in crude form by Thermo Fisher Scientific (Ulm, Germany). HPLC purification yielded peptides with a concentration precision equal to or better than 5%, >97% peptide purity, and >99% isotopic enrichment. Purified peptides were delivered as single-use aliquots in glass tubes.

Trypsinolysis of the UCRP Standards

Each UCRP standard (1.5 nmol) was reduced with 6 mM dithiothreitol (Sigma-Aldrich, St. Louis, MO) in 50 mM ammonium bicarbonate (Sigma-Aldrich) at 50 °C for 45 min and then alkylated with 42 mM iodoacetamide (Sigma-Aldrich) in 50 mM ammonium bicarbonate at room temperature for 45 min. The reduction and alkylation steps can be omitted for the digestion of the pure UCRPs because they do not contain free cysteine residues.

The pH was adjusted to 8.5 with 1 M NaOH (Sigma-Aldrich) and the sample was digested overnight with trypsin (sequencing-grade trypsin, Promega, Madison, WI) using a ratio of 1:20 (w/w) at 37 °C. The digested sample was desalted using Sep-Pak C18 reverse-phase cartridges (Waters, Milford, MA). Peptides were eluted using 1 mL of 70% acetonitrile (CHROMASOLV Plus, Sigma-Aldrich) in 0.1% formic acid (Sigma-Aldrich) and dried using a vacuum concentrator. The dried samples were stored at −20 °C until LC−MS analysis.

External and Direct Calibration in a Single LC−MS run

Ten synthetic SIL peptide variants of the AALPAAFK URP sequence with various combinations of ^{15}N- and ^{13}C-labeled amino acids (labeling underlined) were synthesized and solutions of different concentrations were prepared (2 amol/μL $\underline{\text{AALPAAFK}}$, 4 amol/μL $\underline{\text{AAL}}\text{P}\underline{\text{AAFK}}$, 14 amol/μL $\text{A}\underline{\text{A}}\text{L}\underline{\text{PAAFK}}$, 41 amol/μL $\text{AA}\underline{\text{LPAAF}}\text{K}$, 124 amol/μL $\underline{\text{AALPAAFK}}$, 370 amol/μL $\text{AA}\underline{\text{L}}\text{PA}\underline{\text{AF}}\text{K}$, 1.1 fmol/μL $\text{AALP}\underline{\text{AA}}\text{F}\underline{\text{K}}$, 3.3 fmol/μL $\text{AA}\underline{\text{LP}}\text{A}\underline{\text{A}}\text{FK}$, 10.0 fmol/μL $\text{AAL}\underline{\text{P}}\text{AAFK}$, and 30.0 fmol/μL AALPAAFK). These solutions were spiked: (i) with a UCRP standard previously digested with trypsin for external calibration and (ii) into a biological sample containing an undigested UCRP standard. LC−SRM analysis was performed on a triple quadrupole instrument as described later.

Preparation of Biological Samples

Collection of Urine Samples. Pooled human urine was provided by the Integrated Biobank of Luxembourg (IBBL). Midstream urine samples were collected from 10 nonsmoking healthy volunteers, five females and five males, aged between 30 and 40 years. There was no history of renal dysfunction in any of the subjects, and the individuals were not medicated at the time of sample collection. Urine samples were centrifuged at 1000g for 20 min at room temperature. The supernatants were pooled and stored as 50 mL aliquots in falcon tubes at −80 °C.

Treatment of Urine Samples. The amount of urinary proteins was determined by the pyrogallol assay (Sigma-Aldrich). Samples corresponding to ∼250 μg of urinary protein were precipitated overnight with acetonitrile at a ratio of 1:5 (v/v). After centrifugation at 14 000g for 30 min at 4 °C, the pellets were washed once with acetonitrile, air-dried, and suspended in 250 μL of 8 M urea (Sigma-Aldrich) in 0.1 M ammonium bicarbonate. The samples were reduced with 20 mM dithiothreitol in 50 mM ammonium bicarbonate at 37 °C for 30 min and alkylated with 80 mM iodoacetamide in 50 mM ammonium bicarbonate at 37 °C for 30 min. Sample volumes were adjusted to reach a 2 M urea concentration using 100 mM ammonium bicarbonate. Samples were then digested overnight at 37 °C with trypsin at a 1:20 (w/w) ratio, and digestion was stopped by the addition of formic acid to reach a pH 2. Sep-Pak C18 reverse-phase cartridges were used to clean up and desalt the samples after digestion; the peptides were eluted using 1 mL of 50% acetonitrile in 0.1% formic acid and dried using a vacuum concentrator. The dried samples were stored at −20 °C until the LC−MS analysis.

For the preparation of dilution series, dried urine samples were solubilized in 0.1% formic acid at a final concentration of 1 μg/μL. Five SIL peptides (DGAGDVAFVK, SASDLTWDNLK, EGYYGYTGAFR, LLLTSAPSLATSPAFR, YDLLDLTR) were independently spiked into the urine digest to prepare in parallel two types of samples: first, an urine digest containing calibrated SIL peptides derived from the UCRP standards (0, 0.02, 0.05, 0.1, 0.4, 0.9, 2.3, 5.9, 15.2, and 39.4 fmol/μL) before applying a correction after recalibration based on the response factor, and second, a urine digest containing freshly prepared, conventional, SIL peptides (0, 0.02, 0.05, 0.1, 0.4, 0.9, 2.3, 5.9, 15.2, and 39.4 fmol/μL), with both C-terminal ^{15}N and ^{13}C-labeled arginine and lysine residues (labeling underlined).

Collection and Treatment of Plasma Samples. Blood serum obtained from patients diagnosed with lung cancer was provided by the Integrated Biobank of Luxembourg (IBBL). The two most abundant proteins (human albumin and IgG) were depleted using a multiple affinity removal spin cartridge (MARS 2, Agilent Technologies). Protein concentrations were measured by the Bradford assay (Sigma-Aldrich) before and after the depletion process. After depletion, the sample was supplemented with the first analogue of a UCRP standard acting as a surrogate of human SAA-1 protein, GPGGVWAAEAISDARAALPAAKF (S1−H_A), and then reduced with 10 mM dithiothreitol at 50 °C for 50 min, alkylated with 25 mM iodoacetamide in the dark at RT for 30 min, and digested with trypsin using a ratio of 1:20 (w/w). The sample was supplemented with the URP calibration mixture and the calibrated PRP from the second analogue of the UCRP standard of interest, GPGGVWAAEAISDARAALPAAFK (S1−H_B), and then desalted on C18 cartridges (elution with 0.1% formic acid/50% acetonitrile/water) and dried using a vacuum concentrator. The sample was solubilized in 0.1% formic acid to obtain a final concentration of 1 μg/mL. LC−SRM analysis was performed on a triple quadrupole instrument as described later.

Liquid Chromatography and Mass Spectrometry

LC Separation. All peptide separations were carried out on a Dionex Ultimate 3000 RSLC-nano system (Thermo Scientific). For each analysis, the sample was loaded into a Dionex Acclaim PepMap trap column (2 cm × 75 μm i.d., C18, 3 μm, 100 Å) at 5 μL/min using an aqueous solution of 0.05% (v/v) trifluoroacetic acid (Sigma-Aldrich) and 1% acetonitrile. After 3 min, the trap column was set online with a Dionex Acclaim PepMap RSLC analytical column (15 cm × 75 μm i.d., C18, 2 μm, 100 Å). Peptide separation was performed by applying a mixture of solvent A/B. Solvent A was HPLC-grade water with 0.1% (v/v) formic acid, and solvent B was HPLC-grade acetonitrile with 0.1% (v/v) formic acid. Separations were performed by applying (i) a linear gradient of 2−35% solvent B in solvent A at 300 nL/min over 48 min, followed by a washing step (5 min at 90% solvent B in solvent A) and an equilibration step (10 min at 2% solvent B in solvent A) or (ii) a stepwise gradient of 17% solvent B in solvent A over 5 min, followed by a washing step (4 min at 90% solvent B in solvent A) and an equilibration step (10 min at 2% solvent B in solvent A). Sample injection volume was 1 μL.

Analyses on a Quadrupole-Orbitrap Instrument. SIM and PRM analyses were performed using a Q-Exactive mass spectrometer (Thermo Scientific, Bremen, Germany). A dynamic nanoelectrospray source was utilized with uncoated silica tips of 12 cm length, 360 μm outer diameter, 20 μm inner diameter, and 10 μm tip inner diameter. For ionization, 1500 V of liquid junction voltage and capillary temperature of 250 °C were used. For the analyses of the dilution series of five peptides (calibrated SIL peptides derived from the UCRP standards or conventional SIL peptides) in urine samples (performed in triplicate), the acquisition method combined two scan events corresponding to a full-scan method and a time-scheduled sequential PRM method targeting the five pairs of SIL peptides/endogenous peptides in ±1 min retention time windows. The full-scan method employed a m/z 300−1500 mass selection, an Orbitrap resolution of 70 000 (at m/z 200), a target automatic gain control (AGC) value of 1×10^6, and maximum fill times of 250 ms. The time-scheduled PRM method employed an Orbitrap resolution of 35 000 (at m/z 200), a target AGC value of 1×10^6, and maximum fill times of 120 ms. The precursor ion of each targeted peptide was isolated using a 2 m/z unit window. Fragmentation was performed with a normalized collision energy of 25 eV, and MS/MS scans were acquired with a starting mass of m/z 100, with the ending mass being automatically defined by the m/z and the charge state of the precursor ion. Data analysis was performed using Pinpoint (version 1.2 Thermo Fisher Scientific). Ion chromatograms were extracted with a mass tolerance of 10 ppm for SIM data and 20 ppm for PRM data.

Analyses on a Triple−Quadrupole Instrument. Selected reaction monitoring analyses were performed using a TSQ Vantage extended mass range triple−quadrupole mass spectrometer (Thermo Scientific, San Jose, CA) with identical nanoelectrospray and chromatographic settings as previously described. The selectivity for both Q1 and Q3 quadrupoles was set to 0.7 Da (fwhm). The argon collision gas pressure in the second quadrupole Q2 was set at 1.5 mTorr. For each peptide, the selection of the monitored transitions and the optimization of the collision energy required were performed as described previously.[22] Data analysis was performed using Pinpoint (version 1.2 Thermo Fisher Scientific) or Skyline.

Quantification Based on Isotope Dilution Strategy

For the analyses of dilution series using SRM or PRM methods, the area under the curve (AUC) of each targeted transition (SRM analysis) and selected fragment ion (PRM analysis) was determined for each dilution point. For the analyses of the dilution series of URP isotopologues in a single LC−SRM run, the peptide AUCs were directly used to establish the corresponding dilution curves. For the analyses of the dilution series of the five SIL peptides in urine samples by PRM analysis, the peptide AUCs were employed to calculate SIL/endogenous peptide AUC ratios. These SIL/endogenous peptide AUC ratios were then used to establish the dilution curves of each peptide. For each dilution series, a linear regression analysis was performed. The range of linearity was defined as the range of spiked peptide amounts for which the relative difference between calculated concentrations and the spiked concentrations was <20%. The results were combined with the SIL/endogenous peptide area ratios of three replicate analyses per dilution point with CVs lower than 20% and were used to determine the amount of endogenous peptides in urine samples.

■ RESULTS

Cleavable Reporter Peptide Standard

The proposed concatenated polypeptide standard (UCRP standard), incorporating a tryptic cleavage site, contains the sequence of a signature peptide, or a PRP, concatenated to a URP (Figure 1). Upon digestion, the URP is cleaved off and can be precisely quantified. This result is used to determine the concentration of the released PRP. The concept is based on the stoichiometric release of both PRP and URP during tryptic digestion (equimolar concentration/amount). The URP sequence is universal to all UCRP standards and therefore can be used to recalibrate any of this sequence-specific PRP. The UCRP standards have been designed with a tryptic cleavage site to widespread their application in standard bottom-up proteomic approaches. Trypsin is the most commonly used enzyme in the field of proteomics due its superior efficiency and specificity, although a number of alternative proteases also have been successfully applied.[23−26]

The choice of the URP sequence was based on several criteria. A prerequisite was the nonexistence of the sequence in the UniProt KB database (version 2011_10) to allow the unequivocal determination of its concentration and of the released PRP in a variety of samples from several species. This is especially important for the "calibration of the PRP (internal calibration) and quantification of the corresponding endogenous peptide of interest" in a single LC−MS run. The selection of the URP sequence included good LC−MS detectability (based on amino acid composition, hydrophobicity factor), trypsinization specificity (based on amino acid patterns adjacent to the cleavage site), synthesis constraints with respect to the final length of the UCRP standards, and, to a minor extent, the cost of synthesis (sequence including amino acids that are frequently labeled in SIL peptides were preferentially selected, (i.e., L, V, and A)). Thus, URP sequences consisting of eight amino acids were selected based on experimental LC−MS evidence, including peptide elution profiles, ionization efficiency, and fragmentation pattern. Then, preliminary experiments to assess tryptic proteolysis were performed on a series of UCRP standards created by fusing 16 distinct URP sequences (SI, Table S-1) to a single PRP. Three PRPs were carefully selected based on their amino acid composition as positive and negative controls for proteolysis (taking into account the fusion at the N-or-C-termini) based on the miscleavage peptide patterns established by Thiede et al.[27] The sequences of the three PRPs were SFFSFLGEAFDGAR, ELDESLQVAER, and ASSIIDELFQDR, respectively labeled PRP1, PRP2, and PRP3. The cleavable reporter peptide standards were digested by trypsin at 37 °C for 12 h. A desalting step was performed before injection in the LC−MS system. The efficiency of the enzymatic digestion was found to be optimal by positioning the PRP in front of the URP. This is consistent with previous reports[27,28] indicating that trypsin cleaves less efficiently after lysine residues as compared with arginine residues. Out of those UCRP standards with PRPs in front of the URP, three URP sequences ([AALPAAFK], [AANFAAFK], and [AAQLAALK]) showed the best efficiency of proteolysis for the three combinations of resulting concatenated polypeptides (PRP1-URP, PRP2-URP, and PRP3-URP). More specifically, the proteolysis by trypsin was nearly complete (>99%) for the UCRP standards containing the AALPAAFK URP sequence. From this set of experiments, the position of the URP was fixed C-terminal in the resulting cleavable standards, and the AALPAAFK URP sequence was retained for further investigation.

To better assess the proteolysis efficiency, we performed a second set of experiments on a new set of UCRP standards containing the selected URP sequence: AALPAAFK. It was fused to 43 additional PRPs (PRP4-URP to PRP46-URP) for UCRP standard synthesis (SI, Table S-2). The efficiency of proteolysis obtained with the AALPAAFK URP sequence for the 43 PRPs was determined to be 95.8% (CV 8.9%), demonstrating that the selected URP can be cleaved off the UCRP standards regardless of the PRP sequence.

Finally, the trypsinization efficiency of the UCRP standard was determined using a SIL analogue (identical sequence with different SIL amino acids) of the standard. An example is illustrated in Figure 2. Undigested and trypsin digested samples (Samples 1 and 2, respectively) of the UCRP standard GPGGVWAAEAISDARAALPAAFK were spiked with a defined amount of the undigested SIL analogue GPGGVW-AAEAISDARAALPAAFK (labeling underlined). LC−SRM analysis monitoring the UCRP standard, its isotopologue, and the trypsin digestion products (both PRP and URP) was performed in triplicate (Figure 2). The UCRP standard was found to be very effectively digested by trypsin. The SRM measurements were sufficiently accurate to determine a trypsin digestion rate of >99% as the S/N ratio of the residual undigested standard was lower than 1% of the S/N ratio of the standard prior digestion (Figure 2). Following this methodology, the efficiency of the trypsin digestion of the UCRP standard can be precisely determined.

However, it is important to notice that due to the resulting inherent equimolar amount of both peptides upon enzyme digestion, (i.e., PRP and URP), an unexpected incomplete proteolysis of the UCRP standard is not of relevance for subsequent calibration.

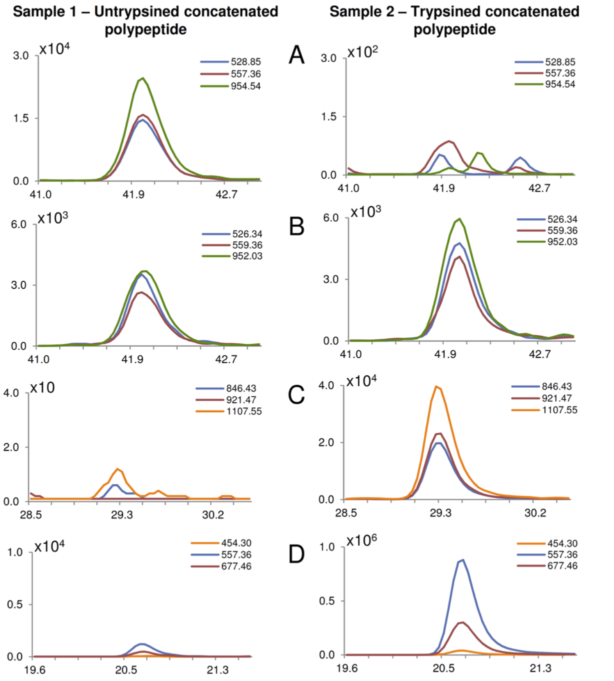


Figure 2. Determination of proteolysis efficiency. The UCRP standard GPGGVWAAEAISDARAALPAAFK (m/z 761.10, z = 3; panel A) is hydrolyzed into the PRP sequence GPGGVWAAEAISDAR (m/z 737.87, z = 2; panel C) and the URP AALPAAFK (m/z 410.27, z=2; panel D). An analogue of the UCRP standard, GPGGVWAAEAISDARAALPAAFK (m/z 757.42, z = 3; panel B), is spiked as internal standard before the LC–MS analysis in both samples: undigested (Sample 1) and digested (Sample 2). Isotopically labeled amino acid residues are underlined.

Polypeptide Standard Application for Protein Quantification

To provide an example of application of the proposed UCRP standards methodology to biological samples, we performed the targeted quantification of two proteins in urine samples. These two proteins, serotransferrin (TRFE) and alpha-N-acetylglucosaminidase (ANAG), have been reported as usually present in urine samples.[29−31] Transferin was reported to be expressed in high level in urine samples from bladder cancer patients, while ANAG is involved in the degradation of heparan sulfate.[30,31]

PRP Selection. Peptides were selected as surrogates for these two proteins for their precise quantification. The peptides DGAGDVAFVK, SASDLTWDNLK, and EGYYGYTGAFR were selected as PRPs for TRFE and LLLTSAPSLATSPAFR and YDLLDLTR as PRPs for ANAG. Five UCRP standards were synthesized by fusing the PRP with C-terminal ^{15}N- and ^{13}C-labeled arginine or lysine residues and the AALPAAFK URP sequence as follows: DGAGDVAFVKAALPAAFK, SASDLTWDNLKAALPAAFK, YDLLDLTRAALPAAFK, EGYYGYTGAFRAALPAAFK, and LLLTSAPSLATSPAFRAALPAAFK (labeling underlined).

Response Factors: Determination of the Relative Response. The relative response factor between the 1:1 stoichiometric released products enables the determination of partial losses of SIL peptides during storage. Consequently, the recalibration of SIL peptides improves the precision/accuracy of the peptide quantification. Conversely, with the relative response factor known, the URP can be used to precisely determine the actual concentration of PRP in the sample. The method can be used to determine peptide concentrations on a routine basis. In addition, such a simple analysis (trypsinolysis of the standard followed by an LC−MS analysis) is straightforward and can be performed at moderate cost and effort.

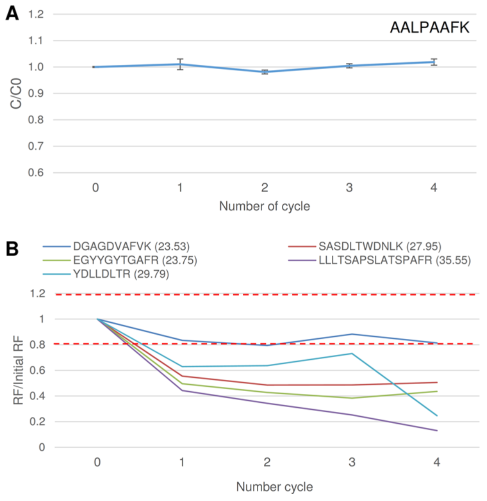


Figure 3. Assessment of peptide recovery upon storage. (A) Recovery of the AALPAAFK URP sequence over four repeated freeze−thaw cycles. Storage was performed at −20 °C. The Y axis represents the peptide concentration (C) after freeze−thaw cycle(s) divided by its initial concentration (C0). (B) Recovery of the PRP (based on the relative response factor determined after digestion (37 °C, pH 8.5, 12h)) repeated after four freeze−thaw cycles. The Y axis represents the signal measured after the freeze−thaw cycle(s) divided by its initial signal. The peptide recovery range between 80 to 120% is delimited by red lines. The relative response factors and the hydrophobicity indexes of the various sequences are indicated in parentheses: DGAGDVAFVKAALPAAFK (1.16, 23.53), SASDLTWDNLKAALPAAFK (0.62, 27.95), EGYYGYTGAFRAALPAAFK (0.93, 23.75), LLLTSAPSLATSPAFRAALPAAFK (0.69, 35.55), and YDLLDLTRAALPAAFK (1.64, 29.79).

The validity of the approach is dependent on the URP sequence characteristics with respect to chemically stability, especially during long-term storage in solution (absence of oxidation and deamidation), and a high recovery rate of the peptides. An analysis after several freeze−thaw cycles showed excellent recovery of the AALPAAFK URP sequence stored in solution using low-adsorption plastic tubes. After four freeze−thaw cycles, the variability was determined to be <2% (Figure 3A). The relative response factors between PRP and URP were established by repeated proteolysis (n = 5) of the standard (Figure 3B). Coefficients of variation of the response factor measurements were all <21% in replicate experiments. The relative response factors, ranging from 0.6 to 1.2, reflect the different ionization efficiencies of each PRP. Peptide recovery was assessed by systematic measurement of the response after repeated freeze−thaw cycles (storage cycles) and the response factors previously determined (Figure 3B). Partial losses of all PRPs released from tryptic digestion were observed after four freeze−thaw cycles, varying between 19 to 87%. Interestingly, the first freeze−thaw cycle appears to be the most critical, accounting for ~50% of the total loss for three out of five PRPs. The peptide showing the poorest recovery upon storage (close to 90% loss) was hydrophobic in nature with a hydrophobicity factor of 35.5 (Figure 3B). Consequently, this peptide may undergo hydrophobic interactions with surfaces affecting the concentration of the peptide in solution. These results show the

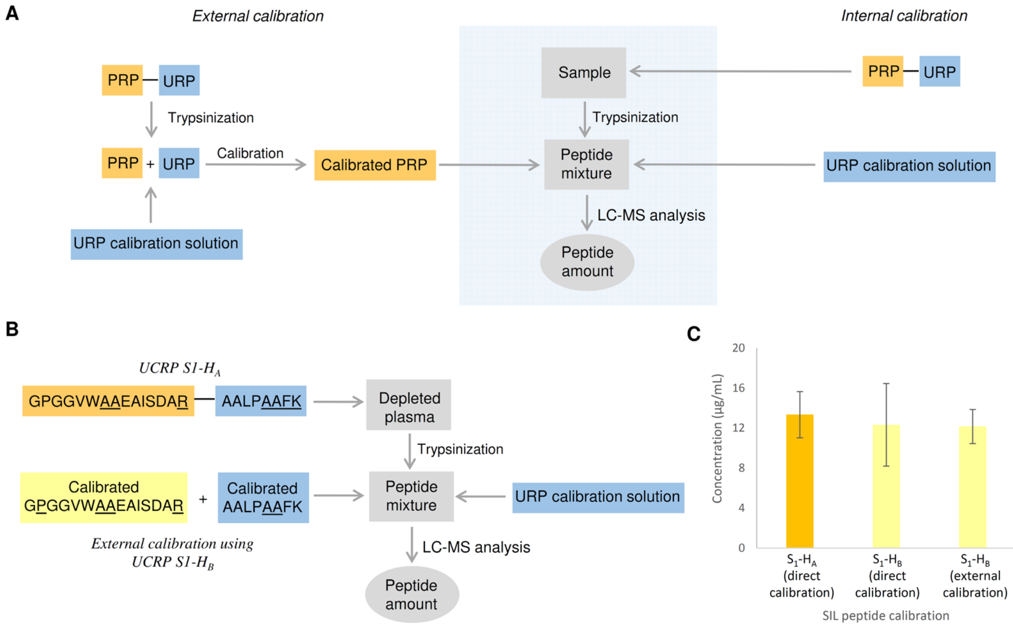


Figure 4. Calibration strategies using UCRP standards. (A) Two options are considered to calibrate a SIL peptide in a single LC−MS run: the external calibration, where the PRP peptide is first calibrated in buffer and then spiked in the sample of interest already digested to estimate the endogenous concentration of the peptide of interest, and the internal calibration, where the concatenated polypeptide standard is spiked in the sample before trypsinization and then the URP isotopologues are spiked in the peptide mixture before LC−MS run analysis to concomitantly calibrate the SIL peptide and estimate the endogenous peptide concentration in a single run. (B) Comparison of peptide quantification in samples using internal and external SIL peptide calibrations. By using two UCRP analogues per peptide of interest, three measurements of the endogenous peptide were performed out of a single LC−MS analysis. (C) Values for the peptide GPGVWAAEAISDAR measured in a pooled serum sample using the different UCRP approaches.

negative effects of storage conditions and freeze−thaw cycles on conventional SIL peptides and address fundamental issues of peptide recovery upon storage, which often have not been taken into account in previous protein quantification studies. Systematic recovery studies on synthetic peptides in relation to storage conditions are laborious and inadequate for a large scale study; consequently, our results indicate the need for recalibration of SIL peptides before usage. The usefulness of the UCRP standard approach described here represents an attractive route to achieve accurate protein quantification results.

URP Calibration in One Single LC−MS Analysis. Two options are available to calibrate the URP released from the standard during trypsinization. The simplest approach is based on a single reference point, with an isotopically labeled analogue of the URP spiked at a known amount into the digested standard sample, and both peak areas (URP and URP analogue) used to estimate the amount of the URP. An alternative approach involves URP quantitation using the reverse curve method. In this latter case, defined amounts of several URP analogues (URP calibration mixture) are spiked into the sample to create a series of concentration standards. The reverse curve approach may be better suited for analyses in complex matrices.[32] While the URP is universal for all UCRP standards synthesized, we have exploited the property of isotopologues (i.e., various isotope labeling on individual amino acids based on $^{15}N/^{13}C$ incorporation) to generate an URP standard curve in a single LC−MS analysis as previously described.[33]

Potential interferences are often observed from coeluting components with near-isobaric fragment ions at similar m/z values. Ten isotopologues with the AALPAAFK URP sequence were designed, synthesized, and purified. A mass difference of 5 Da (2.5 Th for doubly charged precursors) between adjacent isotopologues was chosen to ensure minimal interferences (SI, Table S-3). To avoid potential interferences during the mass selection in SRM experiments, we characterized the products using a hybrid quadrupole-orbitrap instrument operated in SIM mode. The individual URP isotopologue ions were analyzed using narrow windows (SI, Figure S-1 depicts the reconstructed MS spectra derived from the individual SIM measurements of the 10 URP isotopologues, indicating no overlapping isotopic distributions). This allows quantification of the isotopologue mixture in a single LC−MS analysis in a conventional SRM experiment.

The dilution curves constructed from the 10 URP isotopologues clearly demonstrate the linearity of the measurements in buffer, in urine, and in plasma samples (SI, Figure S-2). An additional criterion was used to determine the limit of quantification. The LOQ was established for the lowest concentration, with a CV < 20% and accuracy ranging between 80 and 120%, to be at 5, 40, and 125 amol injected into the column for the reporter in buffer, urine samples (corresponding to 250 amol/μg of total urinary protein), and plasma samples (corresponding to 1.3 fmol/μL of plasma), respectively (SI, Figure S-2). These values indicate that the URP peptide is not affected by ionization suppression. It also reflects that the transitions selected for the SRM assay of the URP are not interfered by the background, indicating the robustness of the assay. In conclusion, the quantification of the universal reporter using the 10 URP isotopologues is both accurate (range: 96−113%) and precise (average: 8.3%) in buffer and in complex matrices (urine and plasma).

Multiplexed PRP Calibration in One Single LC−MS run. The calibration of URPs in one single LC−MS analysis using UCRP standards constitutes a new, effective, rapid, and

Table 1. Quantification of Peptide Surrogates of Serotransferrin (TRFE) and α-*N*-Acetylglucosaminidase (ANAG) in Pooled Urine Samples Using UCRP Standards or SIL Peptides[a]

		UCRP method		SIL method	
signature peptide	protein	conc. [fmol/μL][b]	amount [pmol/μg][c]	conc. [fmol/μL][b]	amount [pmol/μg][c]
DGAGDVAFVK	TRFE	5.99	1.50	6.18	1.55
SASDLTWDNLK	TRFE	5.88	1.47	5.84	1.46
EGYYGYTGAFR	TRFE	5.85	1.46	7.18	1.80
LLLTSAPSLATSPAFR	ANAG	1.34	0.34	1.06	0.27
YDLLDLTR	ANAG	0.94	0.24	0.82	0.21

[a]Calibrated SIL peptides from UCRP standards were calibrated in buffer and spiked into the digested urine sample at various concentrations to estimate the endogenous concentration of the peptides. In parallel, conventional SIL peptides were freshly prepared and spiked at different concentration into an aliquot of the same digested urine sample. The determination of endogenous peptide amounts were performed using dilution curves. [b]Endogenous peptide concentration (fmol/μL). [c]Endogenous peptide concentration (pmol/μg of total urinary proteins).

accurate method to calibrate the SIL peptides amount in a sample (Figure 4A). In addition, the ability to analyze ten AALPAAFK URP isotopologues without cross-interferences allows flexibility in study design, that is, allows the calibration of a single or multiple PRP(s) in one single LC−MS run. Indeed, the number of URP isotopologues used to perform the standard reverse curve in one single LC−MS analysis is dependent on the number of UCRP standards spiked into the same sample. For calibration of a single PRP using a single UCRP standard, the nine isotopologues available can be spiked in various amounts in the polypeptide digest to generate a standard reverse curve. Calibration of two or more PRPs (using two or more UCRP standards) within the same LC−MS run is possible using a lower number of reference points and by synthetizing the UCRP standards with different incorporation of stable isotopes into various amino acid residues of the URP. The current AALPAAFK URP sequence and the 10 proposed isotopologues allow the calibration of up to six PRPs within the same LC−MS run, that is, using 6 UCRP standards with 6 different URP isotopologues spiked with a four point dilution series of URP isotopologues. Obviously, if fewer than six isotopologues are used to build the standard reverse curve, the level of precision of the PRP amount determination will be decreased.

Protein Quantification. The PRPs derived from the corresponding UCRP standards previously described were recalibrated and used in the same manner as standard SIL peptides to precisely quantify the two proteins of interest in pooled urine samples (Figure 4A1). In parallel, conventional SIL peptides were synthesized and independently used for peptide quantification to conduct a comparison with the PRP results.

UCRP standards and conventional SIL peptides were freshly prepared at concentrations similar to the proteins found in urine (~1 μg/μL). Dilution series were independently prepared for the two types of standards (concatenated and SIL peptides). The samples were analyzed in triplicate using a quadrupole-orbitrap instrument operated in time-scheduled PRM mode targeting the five pairs of SIL and endogenous peptides. The full set of standards was unambiguously detected as well as the endogenous peptides. The SIL and endogenous peptide AUC ratios were calculated based on the ion chromatograms of each pair of doubly charged precursor ions and were used to establish the dilution curves. For both concatenated standard derived PRPs and SIL peptides, the corresponding dilution curves demonstrated the linear response of measurements in urine samples (SI, Figure S-3). To ensure reliable determination, the purity of the UCRP standards was determined by HPLC. The peptide quantification measured in PRM mode was both accurate (accuracy between 89.2 and 109.2 and between 80.9 and 113.6 using UCRP standards and SIL peptides, respectively) and precise (analytical precision <10% using UCRP standards and <15% using SIL peptides) in the linearity range. The endogenous peptide amounts were estimated using the dilution curve; the quantitative results are shown in Table 1. It is noteworthy that the amount of endogenous peptides determined from the dilution curves was consistent using both types of standards with <18% difference. Excellent agreement was also observed for the endogenous peptide determination (deviation <12%) using SIL peptides freshly prepared or recalibrated PRPs obtained from UCRP standards. As an example, the peptide DGAGDVAFVK, surrogate of serotransferrin, was estimated at 1.55 and 1.50 pmol/μg of total urinary protein using the standard SIL peptide and UCRP standard, respectively (Table 1). These results illustrate the performance of the method based on UCRP standards and its use for the (re)calibration of SIL peptide reference solutions, facilitating the correction of peptide losses upon storage.

The range of applicability of the calibrated PRP (using an external calibration method, Figure 4A) is similar to that conventional SIL peptides, as exemplified for TRFE and ANAG in urine samples. Both the PRP (derived from UCRP) and the conventional SIL peptides have the same properties and behavior as the endogenous analyte (except the mass). It is therefore compatible with any sample preparation protocol aiming at reducing the sample complexity to enhance the performance of the assays, including peptide immune enrichment.[34−36]

Simultaneous PRP Calibration and Quantification of the Corresponding Endogenous Peptide in a Single LC−MS Run

The calibration of PRP derived from an UCRP standard can also be directly performed in the biological samples (internal calibration) as the unique URP sequence permits its unequivocal quantification in a wide range of samples. This allows to simultaneously perform the PRP calibration (internal calibration) and the quantification of the corresponding endogenous peptide in a single LC−MS run (Figure 4A).

The use of the UCRP standard in complex samples is illustrated by the quantification of serum amyloid A-1 protein (SAA1) in plasma. SAA proteins are apolipoproteins associated with several diseases[37] and, increased levels of SAA were found in serum from patients with various cancer types,[38] including lung cancer for which increased levels of SAA1 and SAA2 were reported.[39] Analyses were performed on a pooled sample of sera collected from patients diagnosed with lung cancer. The signature peptide GPGGVWAAEAISDAR was selected for SAA1, and two analogues of the UCRP standards were synthesized with different SIL amino acids, GPGGVWAAEAISDARAALPAAKF (S1−H$_A$) and GPGGVWAAEAISDARAALPAAFK (S1−H$_B$). This allows us to perform two quantitative measurements simultaneously: first, the determination of PRP concentration using the initial method (external calibration, Figure 4A), and, second, the calibration (internal calibration) and quantification of the corresponding endogenous peptide in the sample of interest in a single LC−MS run (Figure 4A). The addition of the UCRP standards to the sample in two distinct forms (digested and nondigested standards) was performed at two distinct stages of the process. As illustrated in Figure 4, the depleted plasma sample was first supplemented with a polypeptide standard (S1−H$_A$) and then digested. Second, the sample was supplemented with the calibrated PRP derived from the second polypeptide standard (S1−H$_B$ after digestion), desalted, and, finally, supplemented with URP isotopologues and analyzed by LC−MS (Figure 4 B). In this way, the measurements of the endogenous peptide (GPGGVWAAEAISDAR) in plasma digest were based in situ and in the "external" digestion using the URP calibration solution. The endogenous peptide concentrations were similar with larger variability for the direct calibration, likely due to interferences of the matrix (Figure 4C). These results demonstrate that the newly developed UCRP standard can be efficiently used as internal standard in a complex matrix (internal calibration) to determine accurate endogenous peptide concentrations (Figure 4 A). This illustrates the versatility of the UCRP method when using different PRP and URP sequences with different isotopic incorporation.

The method could be further expanded to analyze multiple proteins in a multiplexed manner using up to six UCPR standards, thus allowing the concomitant quantification of up to six peptides of interest under the provision that an adequate isotopic labeling scheme is designed. This may become of relevance in the context of clinical applications (high-throughput) of proteomic biomarkers, where only few proteins are to be monitored.[31,40−43]

For the alternative use of the UCRP standard in complex samples, the limiting factor is the LOQ of the URP in the biological matrix for the calibration of the PRP and therefore for the quantification of the endogenous peptide in cases where the PRP has a strongly positive response factor as compared with the URP. The LOQs for URP were determined at 1.3 fmol/μL in HAS/IgG depleted plasma in this study, and the detectability of proteins is well within the range of detection of protein biomarkers as measured with conventional isotopic dilution methods.[44]

CONCLUSIONS

This study addresses SIL peptide recovery issues of relevance for isotopic dilution mass spectrometry. To ensure accurate quantitative results, we used cleavable reporter peptide standards, composed of a URP and a PRP, which allow for the (re)calibration of SIL peptides to correct for potential peptide loss due to storage conditions. This straightforward approach improves quantification accuracy contributing to a more reliable interpretation of biological or clinical data.

The approach yields excellent quantitative results in complex matrices, demonstrating its benefit for accurate quantification of targeted peptides as an alternative to standard SIL peptides. With this approach, the AAA is required only for the calibration of the URP references (isotopologues). The cost of purified UCRP standard is in the same range of conventional purified SIL peptides. However, if several UCRP standards are used over and over again during large scale experiments, the overall costs to check the calibrated amount of samples will dramatically decrease. In addition, the time-efficiency improvements by using UCRP standards are a consequence of the use of a reverse standard calibration curve established from a single LC−MS analysis, as compared with a conventional calibration curve analysis requiring several MS analyses.

We have, in addition, demonstrated that this approach is well-suited for internal calibration in the sample of interest, facilitating the accurate determination of the endogenous peptide concentration in a single LC−MS run. The current URP sequence allows for the multiplexed quantification of several biomarkers in complex mixtures (up to six proteins), but the ongoing development of extended peptide standard sets will further increase the versatility of the approach and expand the applicability of the method.

ASSOCIATED CONTENT

Supporting Information

Evaluated URP sequences. UCRP standards used for proteolysis assessment and corresponding efficiency of proteolysis (%). URP isotopologues and SRM parameters. Evaluation of URP isotopologue interferences. Dilution calibration curves of 10 URP isotopologues. Linearity range of the dilution curves established by PRM analysis from the dilution series of SIL peptides in urine samples. Chromatography traces of quantitative experiments. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Tel: +352 26970-919. Fax: +352 26970-717. E-mail: bdomon@crp-sante.lu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We are grateful to Joel Louette and Diemut Koenig-Greger, ThermoFisher, for their assistance and the synthesis of the reagents. The Fonds National de la Recherche (FNR) of Luxembourg is acknowledged (PEARL and CORE grants). We thank Drs. Sebastien Gallien, Yeoun Jin Kim, Kevin Demeure, and Jan van Oostrum for helpful discussions. Dr. Guy Berchem and IBBL (Integrated Biobank of Luxembourg) are acknowledged for providing the clinical samples.

ABBREVIATIONS

UCRP, universal cleavable reporter peptide standard; PRP, protein reference peptide; URP, universal reporter peptide; SRM, selected reaction monitoring; SIM, selected ion monitoring; PRM, parallel reaction monitoring; MS, mass spectrometry; MS/MS, tandem mass spectrometry; LC, liquid chromatography; AGC, automatic gain control; LOQ, limit of quantification; LOD, limit of detection; AUC, area under the curve; CV, coefficient of variation; S/N, signal-to-noise ratio

REFERENCES

(1) Marx, V. Targeted proteomics. *Nat. Methods* **2013**, *10* (1), 19–22.

(2) Ong, S. E.; Mann, M. Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.* **2005**, *1* (5), 252–262.

(3) Brun, V.; Masselon, C.; Garin, J.; Dupuis, A. Isotope dilution strategies for absolute quantitative proteomics. *J. Proteomics* **2009**, *72* (5), 740–749.

(4) Gerber, S. A.; Rush, J.; Stemman, O.; Kirschner, M. W.; Gygi, S. P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100* (12), 6940–69455.

(5) Beynon, R. J.; Doherty, M. K.; Pratt, J. M.; Gaskell, S. J. Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. *Nat. Methods* **2005**, *2* (8), 587–589.

(6) Brun, V.; Dupuis, A.; Adrait, A.; Marcellin, M.; Thomas, D.; Court, M.; Vandenesch, F.; Garin, J. Isotope-labeled protein standards: toward absolute quantitative proteomics. *Mol. Cell. Proteomics* **2007**, *6* (12), 2139–2149.

(7) Neubert, H.; Muirhead, D.; Kabir, M.; Grace, C.; Cleton, A.; Arends, R. Sequential protein and peptide immunoaffinity capture for mass spectrometry-based quantification of total human beta-nerve growth factor. *Anal. Chem.* **2013**, *85* (3), 1719–1726.

(8) Hebert, A. S.; Merrill, A. E.; Bailey, D. J.; Still, A. J.; Westphall, M. S.; Strieter, E. R.; Pagliarini, D. J.; Coon, J. J. Neutron-encoded mass signatures for multiplexed proteome quantification. *Nat. Methods* **2013**, *10* (4), 332–334.

(9) Carr, S. A.; Abbatiello, S. E.; Ackermann, B. L.; Borchers, C.; Domon, B.; Deutsch, E. W.; Grant, R. P.; Hoofnagle, A. N.; Huttenhain, R.; Koomen, J. M.; Liebler, D. C.; Liu, T.; Maclean, B.; Mani, D. R.; Mansfield, E.; Neubert, H.; Paulovich, A. G.; Reiter, L.; Vitek, O.; Aebersold, R.; Anderson, L.; Bethem, R.; Blonder, J.; Boja, E.; Botelho, J.; Boyne, M.; Bradshaw, R. A.; Burlingame, A. L.; Chan, D.; Keshishian, H.; Kuhn, E.; Kinsinger, C.; Lee, J. S.; Lee, S. W.; Moritz, R.; Oses-Prieto, J.; Rifai, N.; Ritchie, J.; Rodriguez, H.; Srinivas, P. R.; Townsend, R. R.; Van Eyk, J.; Whiteley, G.; Wiita, A.; Weintraub, S. Targeted Peptide Measurements in Biology and Medicine: Best Practices for Mass Spectrometry-based Assay Development Using a Fit-for-Purpose Approach. *Mol. Cell. Proteomics* **2014**, *13* (3), 907–917.

(10) Addona, T. A.; Shi, X.; Keshishian, H.; Mani, D. R.; Burgess, M.; Gillette, M. A.; Clauser, K. R.; Shen, D.; Lewis, G. D.; Farrell, L. A.; Fifer, M. A.; Sabatine, M. S.; Gerszten, R. E.; Carr, S. A. A pipeline that integrates the discovery and verification of plasma protein biomarkers reveals candidate markers for cardiovascular disease. *Nat. Biotechnol.* **2011**, *29* (7), 635–643.

(11) Keshishian, H.; Addona, T.; Burgess, M.; Mani, D. R.; Shi, X.; Kuhn, E.; Sabatine, M. S.; Gerszten, R. E.; Carr, S. A. Quantification of cardiovascular biomarkers in patient plasma by targeted mass spectrometry and stable isotope dilution. *Mol. Cell. Proteomics* **2009**, *8* (10), 2339–2349.

(12) Kuzyk, M. A.; Smith, D.; Yang, J.; Cross, T. J.; Jackson, A. M.; Hardie, D. B.; Anderson, N. L.; Borchers, C. H. Multiple reaction monitoring-based, multiplexed, absolute quantitation of 45 proteins in human plasma. *Mol. Cell. Proteomics* **2009**, *8* (8), 1860–1877.

(13) Lopez, M. F.; Kuppusamy, R.; Sarracino, D. A.; Prakash, A.; Athanas, M.; Krastins, B.; Rezai, T.; Sutton, J. N.; Peterman, S.; Nicolaides, K. Mass spectrometric discovery and selective reaction monitoring (SRM) of putative protein biomarker candidates in first trimester Trisomy 21 maternal serum. *J. Proteome Res.* **2011**, *10* (1), 133–142.

(14) Addona, T. A.; Abbatiello, S. E.; Schilling, B.; Skates, S. J.; Mani, D. R.; Bunk, D. M.; Spiegelman, C. H.; Zimmerman, L. J.; Ham, A. J.; Keshishian, H.; Hall, S. C.; Allen, S.; Blackman, R. K.; Borchers, C. H.; Buck, C.; Cardasis, H. L.; Cusack, M. P.; Dodder, N. G.; Gibson, B. W.; Held, J. M.; Hiltke, T.; Jackson, A.; Johansen, E. B.; Kinsinger, C. R.; Li, J.; Mesri, M.; Neubert, T. A.; Niles, R. K.; Pulsipher, T. C.; Ransohoff, D.; Rodriguez, H.; Rudnick, P. A.; Smith, D.; Tabb, D. L.; Tegeler, T. J.; Variyath, A. M.; Vega-Montoto, L. J.; Wahlander, A.; Waldemarson, S.; Wang, M.; Whiteaker, J. R.; Zhao, L.; Anderson, N. L.; Fisher, S. J.; Liebler, D. C.; Paulovich, A. G.; Regnier, F. E.; Tempst, P.; Carr, S. A. Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat. Biotechnol.* **2009**, *27* (7), 633–641.

(15) Picotti, P.; Aebersold, R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods* **2012**, *9* (6), 555–566.

(16) Holzmann, J.; Pichler, P.; Madalinski, M.; Kurzbauer, R.; Mechtler, K. Stoichiometry determination of the MP1-p14 complex using a novel and cost-efficient method to produce an equimolar mixture of standard peptides. *Anal. Chem.* **2009**, *81* (24), 10254–10261.

(17) Mirzaei, H.; McBee, J. K.; Watts, J.; Aebersold, R. Comparative evaluation of current peptide production platforms used in absolute quantification in proteomics. *Mol. Cell. Proteomics* **2008**, *7* (4), 813–823.

(18) Horinek, D.; Serr, A.; Geisler, M.; Pirzer, T.; Slotta, U.; Lud, S. Q.; Garrido, J. A.; Scheibel, T.; Hugel, T.; Netz, R. R. Peptide adsorption on a hydrophobic surface results from an interplay of solvation, surface, and intrapeptide forces. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (8), 2842–2847.

(19) Grove, J.; Fozard, J. R.; Mamont, P. S. Assay of alpha-difluoromethylornithine in body fluids and tissues by automatic amino-acid analysis. *J. Chromatogr.* **1981**, *223* (2), 409–416.

(20) Winter, D.; Hung, C. W.; Jaskolla, T. W.; Karas, M.; Lehmann, W. D. Enzyme-cleavable tandem peptides for quantitative studies in MS-based proteomics. *Proteomics* **2012**, *12* (23–24), 3470–3474.

(21) Singh, S.; Springer, M.; Steen, J.; Kirschner, M. W.; Steen, H. FLEXIQuant: a novel tool for the absolute quantification of proteins, and the simultaneous identification and quantification of potentially modified peptides. *J. Proteome Res.* **2009**, *8* (5), 2201–2210.

(22) Kim, Y. J.; Zaidi-Ainouch, Z.; Gallien, S.; Domon, B. Mass spectrometry-based detection and quantification of plasma glycoproteins using selective reaction monitoring. *Nat. Protoc* **2012**, *7* (5), 859–871.

(23) Wisniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **2009**, *6* (5), 359–62.

(24) Raijmakers, R.; Neerincx, P.; Mohammed, S.; Heck, A. J. Cleavage specificities of the brother and sister proteases Lys-C and Lys-N. *Chem. Commun.* **2010**, *46* (46), 8827–8829.

(25) Swaney, D. L.; Wenger, C. D.; Coon, J. J. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J. Proteome Res.* **2010**, *9* (3), 1323–1329.

(26) Barrett, A. J. Classification of peptidases. *Methods Enzymol.* **1994**, *244*, 1–15.

(27) Thiede, B.; Lamer, S.; Mattow, J.; Siejak, F.; Dimmler, C.; Rudel, T.; Jungblut, P. R. Analysis of missed cleavage sites, tryptophan oxidation and N-terminal pyroglutamylation after in-gel tryptic digestion. *Rapid Commun. Mass Spectrom.* **2000**, *14* (6), 496–502.

(28) Glatter, T.; Ludwig, C.; Ahrne, E.; Aebersold, R.; Heck, A. J.; Schmidt, A. Large-scale quantitative assessment of different in-solution protein digestion protocols reveals superior cleavage efficiency of tandem Lys-C/trypsin proteolysis over trypsin digestion. *J. Proteome Res.* **2012**, *11* (11), 5145–5156.

(29) Court, M.; Selevsek, N.; Matondo, M.; Allory, Y.; Garin, J.; Masselon, C. D.; Domon, B. Toward a standardized urine proteome analysis methodology. *Proteomics* **2011**, *11* (6), 1160–1171.

(30) Chen, Y. T.; Chen, C. L.; Chen, H. W.; Chung, T.; Wu, C. C.; Chen, C. D.; Hsu, C. W.; Chen, M. C.; Tsui, K. H.; Chang, P. L.; Chang, Y. S.; Yu, J. S. Discovery of novel bladder cancer biomarkers by comparative urine proteomics using iTRAQ technology. *J. Proteome Res.* **2010**, *9* (11), 5803–5815.

(31) Chen, Y. T.; Chen, H. W.; Domanski, D.; Smith, D. S.; Liang, K. H.; Wu, C. C.; Chen, C. L.; Chung, T.; Chen, M. C.; Chang, Y. S.; Parker, C. E.; Borchers, C. H.; Yu, J. S. Multiplexed quantification of 63 proteins in human urine by multiple reaction monitoring-based

mass spectrometry for discovery of potential bladder cancer biomarkers. *J. Proteomics* **2012**, *75* (12), 3529−3545.

(32) Campbell, J.; Rezai, T.; Prakash, A.; Krastins, B.; Dayon, L.; Ward, M.; Robinson, S.; Lopez, M. Evaluation of absolute peptide quantitation strategies using selected reaction monitoring. *Proteomics* **2011**, *11* (6), 1148−1152.

(33) Gallien, S.; Duriez, E.; Crone, C.; Kellmann, M.; Moehring, T.; Domon, B. Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer. *Mol. Cell. Proteomics* **2012**, *11* (12), 1709−1723.

(34) Keshishian, H.; Addona, T.; Burgess, M.; Kuhn, E.; Carr, S. A. Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution. *Mol. Cell. Proteomics* **2007**, *6* (12), 2212−2229.

(35) Kuhn, E.; Addona, T.; Keshishian, H.; Burgess, M.; Mani, D. R.; Lee, R. T.; Sabatine, M. S.; Gerszten, R. E.; Carr, S. A. Developing multiplexed assays for troponin I and interleukin-33 in plasma by peptide immunoaffinity enrichment and targeted mass spectrometry. *Clin. Chem.* **2009**, *55* (6), 1108−1117.

(36) Whiteaker, J. R.; Lin, C.; Kennedy, J.; Hou, L.; Trute, M.; Sokal, I.; Yan, P.; Schoenherr, R. M.; Zhao, L.; Voytovich, U. J.; Kelly-Spratt, K. S.; Krasnoselsky, A.; Gafken, P. R.; Hogan, J. M.; Jones, L. A.; Wang, P.; Amon, L.; Chodosh, L. A.; Nelson, P. S.; McIntosh, M. W.; Kemp, C. J.; Paulovich, A. G. A targeted proteomics-based pipeline for verification of biomarkers in plasma. *Nat. Biotechnol.* **2011**, *29* (7), 625−634.

(37) Uhlar, C. M.; Whitehead, A. S. Serum amyloid A, the major vertebrate acute-phase reactant. *Eur. J. Biochem.* **1999**, *265* (2), 501−523.

(38) Biran, H.; Friedman, N.; Neumann, L.; Pras, M.; Shainkin-Kestenbaum, R. Serum amyloid A (SAA) variations in patients with cancer: correlation with disease activity, stage, primary site, and prognosis. *J. Clin. Pathol.* **1986**, *39* (7), 794−797.

(39) Sung, H. J.; Ahn, J. M.; Yoon, Y. H.; Rhim, T. Y.; Park, C. S.; Park, J. Y.; Lee, S. Y.; Kim, J. W.; Cho, J. Y. Identification and validation of SAA as a potential lung cancer biomarker and its involvement in metastatic pathogenesis of lung cancer. *J. Proteome Res.* **2011**, *10* (3), 1383−1395.

(40) Drabovich, A. P.; Diamandis, E. P. Combinatorial peptide libraries facilitate development of multiple reaction monitoring assays for low-abundance proteins. *J. Proteome Res.* **2010**, *9* (3), 1236−1245.

(41) Tan, G. S.; Lim, K. H.; Tan, H. T.; Khoo, M. L.; Tan, S. H.; Toh, H. C.; Ching Ming Chung, M. Novel Proteomic Biomarker Panel for Prediction of Aggressive Metastatic Hepatocellular Carcinoma Relapse in Surgically Resectable Patients. *J. Proteome Res.* **2014**, *13* (11), 4833−4846.

(42) Cohen, A.; Wang, E.; Chisholm, K. A.; Kostyleva, R.; O'Connor-McCourt, M.; Pinto, D. M. A mass spectrometry-based plasma protein panel targeting the tumor microenvironment in patients with breast cancer. *J. Proteomics* **2013**, *81*, 135−147.

(43) Rosenzweig, C. N.; Zhang, Z.; Sun, X.; Sokoll, L. J.; Osborne, K.; Partin, A. W.; Chan, D. W. Predicting prostate cancer biochemical recurrence using a panel of serum proteomic biomarkers. *J. Urol.* **2009**, *181* (3), 1407−1414.

(44) Huttenhain, R.; Soste, M.; Selevsek, N.; Rost, H.; Sethi, A.; Carapito, C.; Farrah, T.; Deutsch, E. W.; Kusebauch, U.; Moritz, R. L.; Nimeus-Malmstrom, E.; Rinner, O.; Aebersold, R. Reproducible quantification of cancer-associated proteins in body fluids using targeted proteomics. *Sci. Transl. Med.* **2012**, *4* (142), 142ra94.

Annex 4

Evaluation of alternative enzymes to trypsin for enhancing proteomics analysis

S. Trévisiol, D. Ayoub, A. Lesur, S. Gallien, B. Domon

Journal of Proteomics; *submitted July 2015*

Manuscript Title:

Evaluation of alternative enzymes to trypsin for enhancing proteomics analysis.

Authors:

Stéphane Trévisiol, Daniel Ayoub, Antoine Lesur, Sébastien Gallien, Bruno Domon*

Affiliation:

Luxembourg Clinical Proteomics Center (LCP), Luxembourg Institute of Health, Strassen, Luxembourg

***Correspondence should be addressed to:**

Bruno Domon

Luxembourg Clinical Proteomics Center (LCP), Luxembourg Institute of Health, Strassen, Luxembourg

Phone : +352 26970-919

Fax : +352 26970-717

Email : bruno.domon@lih.lu

Abstract:

Most mass spectrometry-based proteomics approaches are centered on bottom-up proteomics, which largely rely on trypsin digestion. Even though it produces peptides that are suitable for mass spectrometry and that provide the highest sequence coverage, a significant part of the proteome sequences is not accessible using trypsin. This can be attributed to peptide length due to the uneven distribution of cleavage sites as well as the intrinsic physico-chemical properties of peptides that render them poorly detectable by mass spectrometry. With the adoption of targeted proteomics approaches, there is an increasing need for protein specific peptides to be used as surrogates for protein quantification. Some proteins do not provide such peptides upon trypsin digestion. Post-translational modifications (PTM) add to this challenge as a peptide containing PTM needs to be generated and detected in order to study the modification. Here, the contribution of other enzymes to increase sequence coverage and the number of possible target peptides was evaluated. This encompassed the assessment of the impact of using these enzymes on the complexity of the digested sample, their complementarity to trypsin, as well as the effect of the optimization of collision energy and its effects on sensitivity in targeted proteomics quantification.

Keywords:

Non-tryptic peptides, parallel reaction monitoring, proteases, targeted proteomics, sequence coverage.

Introduction

Through the last two decades, mass spectrometry has driven the advancement of proteomics. Two major approaches emerged throughout the development of MS-based proteomics: discovery and targeted approaches (1). Discovery proteomics experiments are commonly carried out using a shotgun method, which is based on data dependent acquisition (DDA) (2, 3). It has long been employed in early stage of biomarker discovery and comparative studies. It allows the identification and the quantification of a large number of proteins (up to 10000 in recent studies (4, 5)) in complex biological samples. However, the heuristic nature of ion sampling affects reproducibility (6-8) and generally introduces biases towards abundant proteins. These biases are exacerbated due to the enormous dynamic range and complexity of biological samples that exceed the peak capacity and the sampling rates of LC-MS platforms. Data independent acquisition (DIA) is a more recent discovery approach that consists in fragmenting all ions, thus generating a comprehensive product ions map. It can be performed using sequential isolation windows (typically 10-50 Th) (9, 10) or with no isolation window (11-13) to generate the complex fragmentation spectra. However, the co-fragmentation of precursor ions leads to mixed product ions spectra, challenging data processing and affecting selectivity (14, 15). Elution profiles examination and spectral matching using a reference spectral library are generally used to link precursors to their product ions. Quantification is performed based on peak integration of the extracted fragment ion traces. Acquiring a product ion map of all present peptide ions allows reexamination of data in a targeted way using a predefined set of peptides for which specific spectra and traces can be extracted.

On the other hand, targeted quantification approaches such as selected reaction monitoring (SRM)(16) and the more recent parallel reaction monitoring (PRM)(17) allow for deeper penetration into the proteome with greater sensitivity. Targeted proteomics is generally used for precise and accurate protein quantification but also in experiments aiming at the measurement of a larger number of targeted peptides with less emphasis on the performance of quantification. The latter has significantly increased the proteome coverage of targeted experiment (18), but still performs less than shotgun proteomics. Targeted protein quantification allows for higher sensitivity, wider dynamic range and greater reproducibility of measurements. Targeted proteomics is hypothesis driven; it relies on previous knowledge of the targeted peptides to be analyzed as surrogates for the proteins of interest. When this knowledge is not available, prediction algorithms are used to infer peptide quantotypicity (i.e. uniqueness, flyability, and overall behavior in LC-MS/MS).

All these techniques are bottom-up and based on the digestion of proteins using proteases to produce peptides that are analyzed using an LC-MS platform. Proteolytic cleavage at basic amino-acid residues is the most widely used approach to generate peptides in bottom-up proteomics. The three basic amino-acid residues present in proteomes are lysine, arginine and histidine. To date, no protease is known to cleave at histidine residues. When the histidine residue follows a threonine or a serine, cleavage at histidine can be performed with low specificity using copper III. The cleavage at other histidine sites is 10-100 folds slower (19). Cleavage at lysine and arginine residues can be achieved with several enzymes. Trypsin cleaves specifically at the C-terminus of both residues while Lys-C and Arg-C are specific to lysine and arginine residues, respectively. Alternatively, Lys-N has been described more recently and cleaves at the N-terminus of lysine residues. Generating peptides with a basic residue at their C-termini generally increases ionization efficiency in proton adduct electrospray ionization. The peptide would therefore include two basic groups, the N-terminal α-amino group and the guanidine group (Arg) or ε-amino group (Lys) at its C-terminus, and therefore at least two protonation sites. Protein amino acid composition varies depending on their function and localization. For example, lysine occurrence ranges from 6 to 8 % in extracellular, nuclear and cytoplasmique proteins while it goes down to 4.4 % in membrane proteins. Arginine mean occurrence ranges from 4 to 5 % in all protein classes except for nucleus proteins where it reaches 8.7 % (20).

Alternative to basic amino acid residues cleavage, enzymes like Glu-C and Asp-N cleave at acidic residues. Glu-C cleaves preferably to the C-terminus of glutamic acid; however cleavage after aspartic acid also occurs. Conversely, Asp-N presents cleavage preference to the N-terminus of Asp with fewer occurrences for cleavage at Glu. Glu and Asp mean occurrences are around 5.5 and 5 % respectively (21).

Trypsin remains the most appropriate enzyme providing peptides suitable for mass spectrometry that allow the highest sequence coverage and the largest number of identifications in a complex biological sample. However, some parts of protein sequences are not accessible to trypsin due to an uneven distribution of their cleavage sites (lysine and arginine). These missing sequences can contain important information such as PTMs, mutations etc. Furthermore, in the context of targeted quantification, trypsin does not offer in many cases a sufficient number of targetable proteotypic peptides to distinguish some given isoforms. Other enzymes have been described and are generally used in a complementary fashion to access more comprehensively protein sequences. Recently, Guo et al. described the use of various enzymes and multiple enzyme digestions (48 different independent digestions) to increase the sequence coverage of the HeLa proteome. To estimate the total sequence coverage and digestion complementarity, they measured the PAAC “Proteome Amino Acid Coverage”. While the combination of multiple digestions did not increase significantly the number of protein groups identified, it increased the PAAC by three folds compared to the sole use of trypsin. They also showed that in some cases, non-tryptic peptides may yield better response in SRM experiments, allowing better sensitivity (22).

In this study we report an investigation of the use of alternative or complementary enzymes to trypsin for targeted proteomics experiments. We have evaluated experimentally, and using bioinformatics, the impacts of these enzymes on digestion complementarity, availability of proteotypic peptides and digest complexity. We also evaluated the effects of the collision energy and its impact on tandem mass spectra pattern.

Materiel and methods

Chemicals and reagents

All the chemicals and reagents used were purchased from Sigma-Aldrich (Saint Louis, MO, USA) except for RapiGest surfactant which was purchased from Waters (Manchester, UK). The enzymes trypsin, Arg-C, Asp-N were provided by Promega (Madison, WI USA), Lys-C by Pierce (Carlsbad, CA, USA) and Glu-C by Worthington (Lakewood, NJ, USA). Eight recombinant proteins identified as biomarker candidates for non-small cell lung cancer (NSCLC) (23) were provided by Sigma-Aldrich for osteopontin, by Prospec (East Brunswick, NJ, USA) for endoplasmin, glucose-6-phosphatase dehydrogenase and transaldolase, by Novus Biologicals (Littleton, CO, USA) for alpha-actinin 1, filamin A and zyxin and by Abcam (Cambridge, UK) for lactate dehydrogenase. A total of 159 unlabeled and stable isotope labeled synthetic peptides derived from the eight proteins were obtained from Thermo Fisher Scientific (Ulm, Germany). The stable isotope labeling was performed using ^{15}N and ^{13}C isotopes on various amino acid residues. The plasma used was a pool of human plasma from healthy donors which was obtained from the Integrated BioBank of Luxembourg (IBBL).

Enzymatic digestions of depleted plasma

Plasma depletion was performed as described previously (23). Depleted plasma sample was first maintained during 10 min at 99°C. After cooling to room temperature and addition of 0.1% Rapigest in 50mM ammonium bicarbonate and 10% acetonitrile, cysteines were reduced using 10 mM dithiothreitol (DTT) at 50°C, 50 min followed by alkylation of thiol groups using 25 mM iodoacetamide (IAA) for 50 min at room temperature. The alkylation reaction was then quenched using 3 mM DTT. The plasma sample was then split in two equal parts and digested in parallel in Rapigest 0.1% with trypsin and with Lys-C at pH 8, 12h at 37°C with an enzyme to protein ratio of 1:20. After proteolysis, each sample was acidified with 10% formic acid (Sigma-Aldrich) to reach pH 2-3 in order to precipitate Rapigest.

Enzymatic digestions of eight recombinant proteins and the UPS1 mixture

The eight recombinant NSCLC biomarker candidates were digested individually with five different enzymes in parallel. A total amount of 1.2 µg of UPS1 (Sigma-Aldrich) was digested in parallel with three different enzymes trypsin, Lys-C and Glu-C.

For trypsin, Lys-C, Asp-N and Arg-C digestions, cysteines were reduced using 26 mM DTT (in 8 M Urea in 50 mM ammonium bicarbonate, pH = 8) at 50 °C for 45 mins, then alkylated using 88 mM IAA for 45 min at RT in the dark (except for Arg-C where no alkylation was performed). Digestions were performed at 37 °C for 12 hours with a 1:20 enzyme to protein ratio at 1 M urea. Prior to Arg-C addition, $CaCl_2$ and DTT were added to a concentration of 7.3 mM and 3.4 mM, respectively.

For Glu-C digestion, samples were denatured with 8 M urea in 50 mM phosphate buffer pH 7.8, reduced 30 min with 10 mM DTT at 50°C, alkylated 30 min with 20 mM IAA in the dark at room temperature and the excess of alkylating agent was quenched with 8 mM DTT at room temperature. A first enzymatic digestion using an enzyme to protein ratio of 1:20 was performed in 1 M urea at pH 7.8 during 4h at 37°C followed by a second proteolysis with the same ratio during 12h. After digestion, peptides were desalted using solid phase extraction (SPE) cartridges (Sep-Pak Vac C_{18} 100mg, Waters) and dried. Samples were stored at -20°C before LC-MS analyzes.

Liquid chromatography and mass spectrometry

Liquid chromatography separations were performed on an Ultimate 3000 RSLCnano HPLC system (Thermo Fisher Scientific, San Diego, CA, USA). In all experiments, samples were loaded at 5 µL/min with a 1 % acetonitrile + 0.05 % trifluoroacetic acid solution on a trap column (Acclaim PepMap 2 cm × 75 µm i.d., C18, 3 µm, 100 A; Thermo Fisher Scientific) during 3 min. The separations were performed at a 0.3 µL/min flow rate on an analytical column Acclaim PepMap RSLC 15 cm × 75 µm i.d., C18, 2 µm, 100 A (Thermo Fisher Scientific) using a linear gradient of solvent B into solvent A. Solvent A consisted of water + 0.1 % formic acid while solvent B consisted of acetonitrile + 0.1 % formic acid. The gradient went from 2 % to 35 % B in 66 minutes followed by a 4 minutes plateau at 90% before an equilibration step at 2 % B for 9 minutes.

The LC system was coupled to a quadrupole-orbitrap mass spectrometer (Q-Exactive Plus, Thermo Fisher Scientific). Resolutions are defined at 200 m/z. The LC-MS analyses of plasma digests to generate heat ion maps were performed at 140k resolution. The DDA LC-MS/MS analyses were performed at 70k for the survey scan in the 300 to 1500 m/z scan range and 17.5k for the top 15 MS2 scans with an AGC of 1 million (for both MS1 and MS2) and a maximum injection time of 250 ms and 60 ms for MS1 and MS2 respectively. The normalized collision energy (nCE) was set to 25.

The Parallel reaction monitoring (PRM) experiments consisted of a full scan event and several time-scheduled PRM scan events for the different targeted precursor ion. The full scan was performed with a 17.5k resolution in the 300 to 1500 m/z scan range, 1e6 AGC target and maximum fill times of 50 ms.

PRM events were acquired at the resolution 35k, 1e6 AGC target, 2 m/z isolation window and a maximum fill times of 120 ms. To construct breakdown curves of the 159 synthetic peptides, MS/MS spectra were acquired using six nCE values: 10, 15, 20, 25, 30 and 35 and three minutes chromatographic monitoring windows. A total of 20 inclusion list methods were designed to have a maximum of three precursor ions acquired in the same cycle. Several precursor ions under different charge states were targeted for each peptide, depending on the number of basic amino acids present in the sequence.

Data processing

Peptide identifications and Mascot ion scores determinations were obtained by querying two restricted databases containing for the first one the sequences of the eight lung cancer biomarker candidates and for the second one the sequences of the 48 proteins of the UPS1 protein mixture standard.

To build the breakdown curves of the different precursor ions monitored (based on the number of free amino groups of the corresponding peptides) for the 159 synthetic peptides, the areas of the extracted ion chromatograms of all possible b- and y-type fragment ions (excluding b_1 and y_1 ions) were determined and plotted against the collision energy using an in-house developed tool.

The LC-MS heat maps of depleted human plasma digested with Lys-C and trypsin were generated using Xcalibur (2D maps) and MSight (24) (3D maps).

Biostatistic data were generated after the *in silico* digestion of the NeXtProt database (version 2014-05-27), which contains 20126 protein entries (only canonical forms without missed-cleavages were considered), with various enzymes using in-house developed tools. Hydrophobicity factors of the resulting peptides were calculated using SSRCalc (25).

Results and discussion

To study the complementarity of digestions alternative to trypsin and compare their effects on digest complexity, different model samples were used such as a human plasma sample, a mixture of recombinant protein biomarkers candidates, and a commercial standard proteomics mixture. To compare the behavior of non-tryptic peptides upon CID/HCD fragmentation, breakdown curves were constructed using a set of synthetic peptides which allows an insight into the effects of collision energy on MS/MS fragmentation patterns (complexity and efficiency) in a targeted proteomics context.

i) Orthogonality

Trypsin remains the enzyme of choice in shotgun proteomics as it allows the largest coverage of the proteome as compared with other enzymes. However, as stated earlier, the trypsin non-accessible part of the proteome is significant due to the uneven distribution of the trypsin cleavage sites (lysine and arginine residues), across protein sequences resulting either in very short or very long peptides unsuitable for conventional LC-MS analysis. More specifically, peptides shorter than 5-7 amino acids are mainly redundant and nonspecific while peptides larger than 5 kDa tend to have an adverse behavior in classical LC-MS settings. Thus, for targeted protein quantification, protein specific peptides in the 8 – 25 amino-acids residues range are generally selected (26). However, the trypsin inaccessible sequences can be of great interest, especially in cases where particular isoforms or PTMs are of biological significance. In these scenarios, enzymes alternative to trypsin are better suited to access those sequences, provided they produce appropriate peptides.

In order to estimate the orthogonality of alternative enzymes, an *in silico* digestion of the whole human proteome (NeXtProt version 2014-05-27) using trypsin and a set of the common enzymes (Lys-C, Lys-N, Asp-N, Arg-C and Glu-C) was performed. Table 1 compares the number of peptides obtained by several enzymes in the [8 – 25] residues range. The same calculations for peptides in the [5 residues – 5 kDa] range can be viewed in Table SI1.

Table 1: *Distribution of peptides sizes obtained by in silico digestion of the human proteome.*

		Number of peptides					For peptides in the 8 – 25 residues range			
Enzyme	Specificity	Total	< 8 aa non-redundant	>25 aa non-redundant	≥ 8 aa and ≤ 25 aa non-redundant	≥ 8 aa and ≤ 25 aa unique	Average MM (Da)	Median MM (Da)	Average number of residues	Median number of residues
Trypsin	After K/R	1226257	184281	81889	393636	381220	1522.73	1399.57	13.63	12
Lys-C	After K	664321	85953	129267	238738	231273	1658.19	1559.76	14.56	14
Arg-C	After R	656114	81143	140086	239922	233188	1670.14	1577.87	14.78	14
Lys-N	Before K	665812	84687	129182	239000	231531	1658.31	1559.68	14.57	14
Asp-N	Before D/E	1354843	208326	66718	417352	404172	1499.71	1369.65	13.35	12
Glu-C (E/D)	After E/D	1352895	109412	66795	417073	403908	1499.85	1368.85	13.35	12
Glu-C (E)	After E	819699	210149	122038	293168	284940	1635.99	1532.69	14.47	14

Peptides shorter than eight residues are mostly redundant regardless of the enzyme. Roughly, only 20 % (most of them including 5 to 7 residues) are protein specific. However, trypsin and Asp-N produce twice more of these short uninformative peptides than enzymes that cleave only at one residue (Lys-C, Arg-C). Glu-C has a particular behavior as its specificity depends on digestion conditions. As Glu-C cleaves at a slower rate after D and Asp-N cleaves at a slower rate before E (27), partial cleavage is more often observed with these two enzymes.

The primary interest of proteases alternative to trypsin is the accessibility of otherwise unreachable sequences of the proteome. The Lys-C/N and the Arg-C are sub-variants of trypsin activity as they cleave specifically at lysine or arginine, respectively. These enzymes can typically provide access to sequences rich with R or K, otherwise segmented in small peptides when digested by trypsin. On the other hand, Glu-C and Asp-N proteases bring new levels of orthogonality as they cleave at different amino acids residues. To estimate the capacity of these two categories of enzymes to access distinct areas of the proteome from trypsin, the sequence coverage of the human proteome using peptides in the 8 – 25 residues range and three different enzymes was modeled (Figure 1-A).

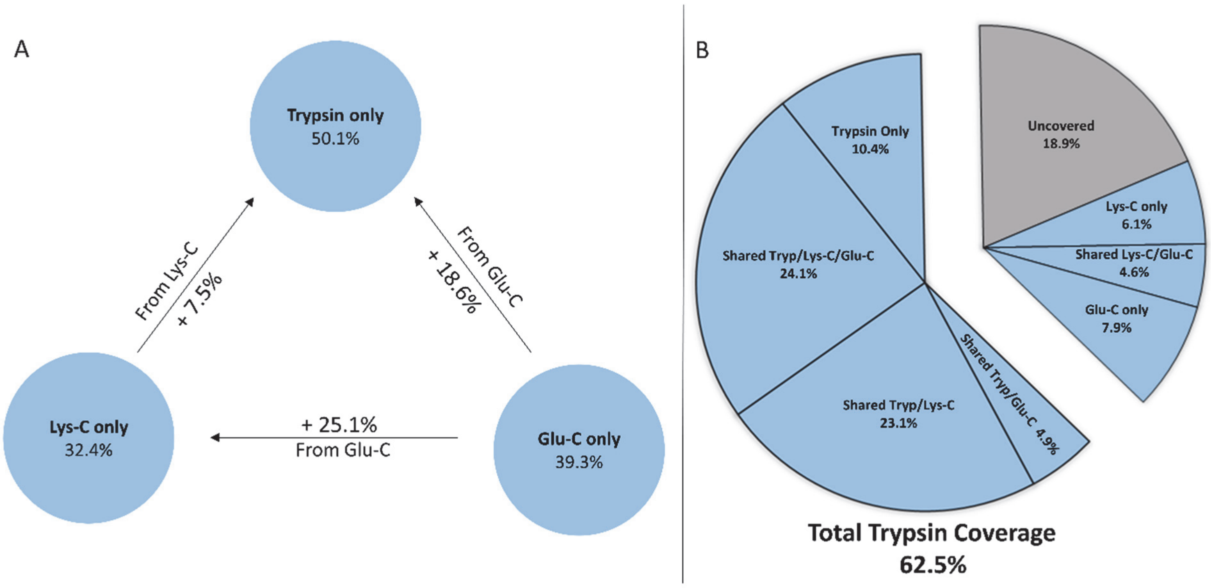


Figure 1: Amino acid coverages of the human proteome with peptides in the 8-25 residues range using multiple *in silico* digestions (A), experimental amino acid coverage of UPS proteins, using trypsin, Lys-C and Glu-C (B).

The simulation shows that Glu-C digestion adds 18.6 % sequence coverage to that of trypsin which corresponds to 37 % of the amino acid sequences originally not accessible by standard tryptic digestion (The 49.9 % missed by trypsin), whereas Lys-C can potentially add 7.5 % of sequence coverage. This observation indicates a degree of orthogonality between Glu-C and trypsin higher than that of Lys-C and trypsin.

To experimentally evaluate the orthogonality of alternative enzymes in accessing different parts of protein sequences, an equal amount of a standard equimolar protein mixture (UPS1) containing 48 human proteins was digested in parallel with trypsin, Lys-C and Glu-C prior to analysis by LC-MS/MS in a regular DDA top 15 experiment. Searches were performed against a restricted database containing solely the UPS1 proteins. Only peptides without any missed-cleavages were used to determine protein coverages. The pie chart (Figure 1-B) represents the overall protein coverage obtained for the standard protein mixture digested with trypsin, Lys-C and Glu-C (considering only the cleavage after glutamic acid residues). As expected, the highest protein coverage is obtained with the tryptic digestion which allowed the identification of 62.5% of the UPS1 proteins. With Lys-C and Glu-C digestions, 57.9% and 41.54% of sequence coverage were achieved, respectively. The lower proteome coverage obtained with Lys-C and Glu-C is consistent with the simulation and other studies (22). Moreover the search engines for peptide identification are generally optimized for tryptic peptides and non-tryptic peptides tend to have lower scores. By merging all the identification information obtained for the three enzymes, the global protein coverage of the UPS mixture rose to 81.09%. In this case, the use of alternative enzymes enabled to recover 49.5 % of the sequence parts that were not covered by trypsin. The propensity of Lys-C and Glu-C to reveal parts of the proteome inaccessible by trypsinization can be exploited to characterize isoforms or mutated sequences. For instance, Lesur et al. have employed a Glu-C digestion to characterize at peptide level the EGF receptor's 746-750 deletion mutation (28). In this particular case, tryptic proteolysis did not generate acceptable signature peptides for the unambiguous characterization of the mutation.

ii) Complexity reduction and impact on LC-MS density

As shown in tables 1 and SI1, trypsin generates nearly twice more peptides than Lys-C, Lys-N or Arg-C and a significant number of those is shorter than 5 amino acid residues. Lys-C/N and Arg-C peptides are 1.45 and 1.51 times larger on average than tryptic peptides, respectively. Generating a smaller number of larger peptides can have beneficial effects in LC-MS analysis. The lower sample complexity is expected to translate in a better separation of the digest components or conversely to allow for the usage of shorter/faster gradients. Moreover, a less complex digest may lead to a reduction in the occurrence of interference in targeted analysis hence an increase in quantification accuracy. The hydrophobicity indexes of the three different proteome digests were calculated and a bar chart of the number of peptides observed in different hydrophobicity index bins was produced (Figure 2). Trypsin, which generates the largest number of peptides, exhibits in the [0-38] hydrophobicity factor range twice more peptides than Arg-C and Lys-C. Interestingly, the Lys-C and Arg-C digests are not richer with hydrophobic peptides compared to trypsin, as it is generally thought, as they have comparable numbers of peptides in the highly hydrophobic region.

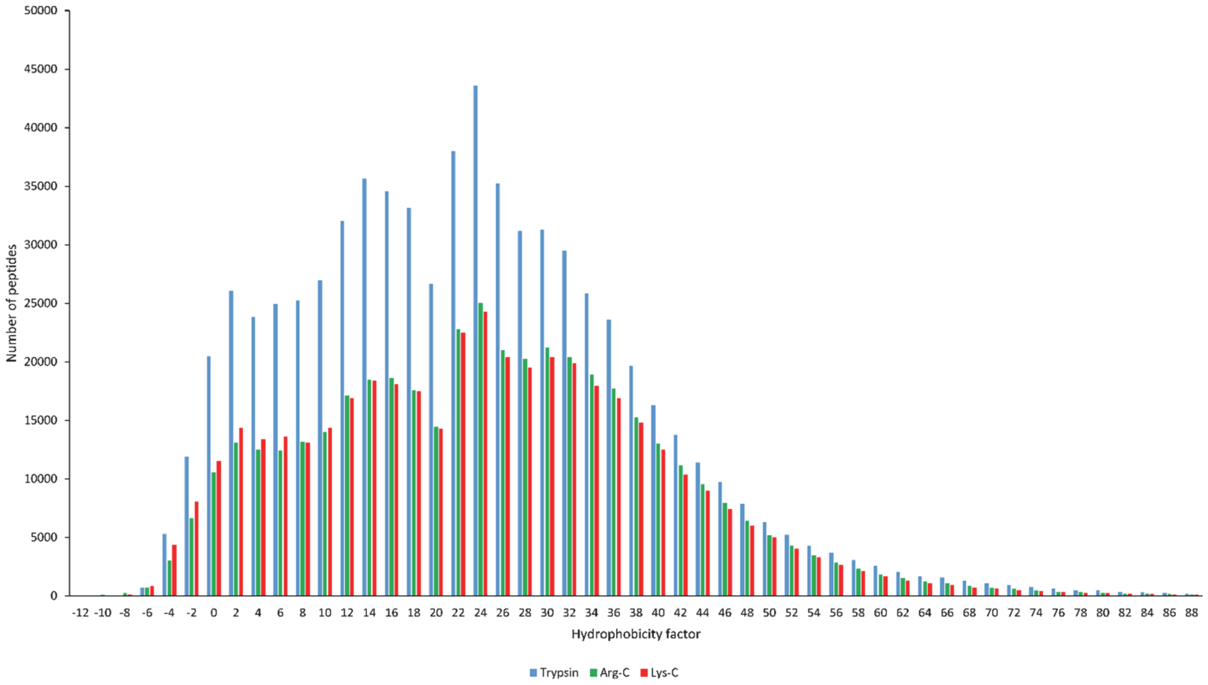


Figure 2: SSRcalc hydrophobicity factor of all peptides in the [5 residues- 5 kDa] range generated by trypsin, Lys-C and Arg-C *in silico* digestion of the human proteome.

In order to further assess the simulation results, a depleted human plasma sample was digested with trypsin and Lys-C and the digests were analyzed with LC-MS using the same gradient. The left panels of Figure 3 present the heat maps of the intensity of measured ions across the chromatographic separation and the m/z range for the two digests. As expected, the trypsin digest occupies more densely the space, especially in the 10-50 minutes range. Similarly to the results obtained by hydrophobicity indexes calculations, the Lys-C digest, which contains larger peptides on average compared to trypsin, do not present a denser area at the end of the gradients which suggests that Lys-C does not produce more hydrophobic peptides than trypsin. Figure 3 right panels represent a three dimensional visualization for the two ranges delimited by the dashed rectangles in the heat maps of depleted plasma (left panels).The reduced number of species in the Lys-C digest decreases the number of ions that fall in the isolation window of quadrupoles when targeting a specific peptide which may result in decreased signal interferences due to co-isolation.

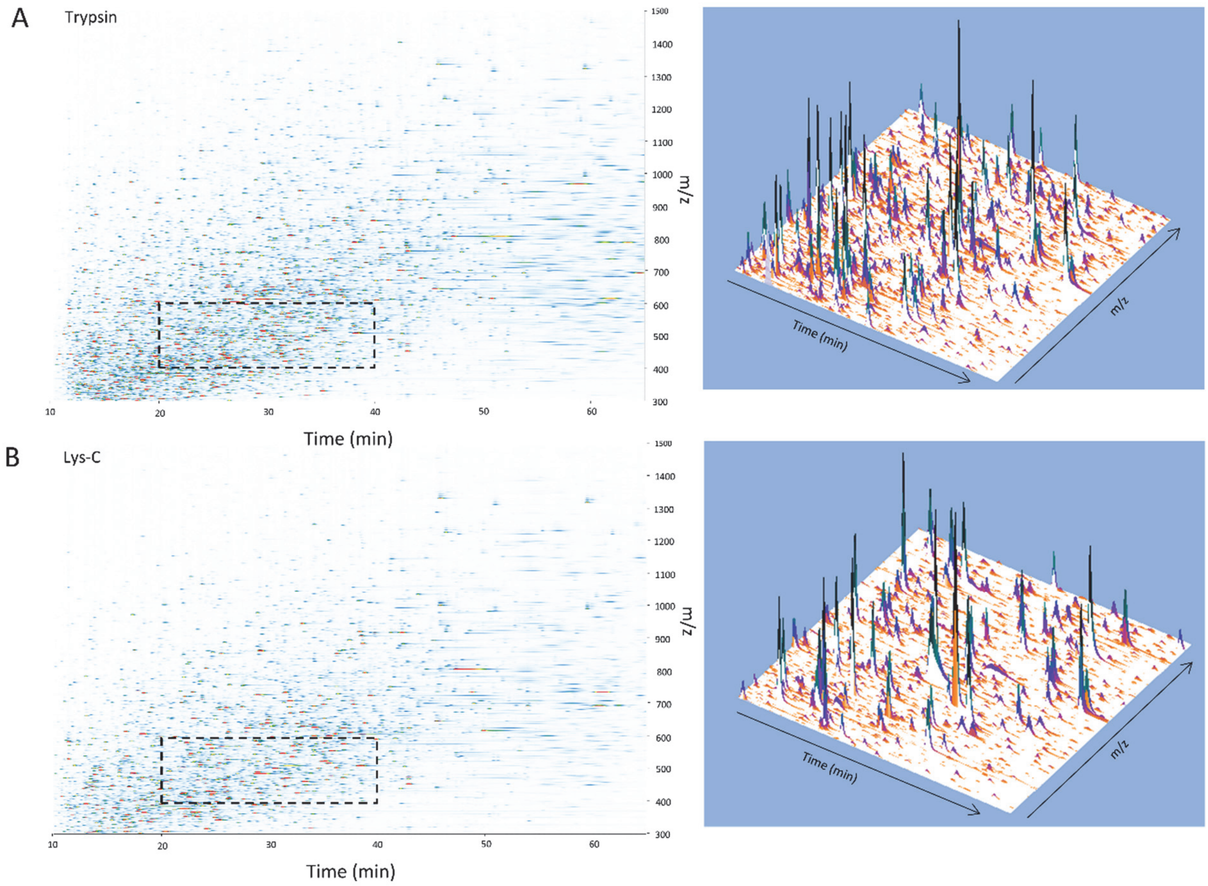


Figure 3: LC-MS heat map of depleted human plasma digested with trypsin (A) and Lys-C (B). The right panels are 3D representations of the peak density in the rectangles delimited by the dashed lines in the left panels.

iii) Collision energies, fragmentation and complexity of MS2 spectra

Non-tryptic peptides often contain internal basic amino acids which influence their MS properties, including their ionization (leading to higher charge states), and their fragmentation patterns. Previous studies (16) aiming at the evaluation of the influence of the parameters affecting CID fragmentation of peptides in the collision cell (in triple quadrupole mass spectrometers or in the HCD cell of orbitrap instruments) showed that collision energy was the main driving factor. The collision energy value, generating the highest intensities of fragment ions, is related to the peptide sequence (including the presence of amino acids promoting facile cleavages such as the proline residues, the number of basic amino acid residues, or the charge state of the precursor ion). The effect of the collision energy on the fragmentation pattern of peptides is best evaluated by generating pseudo-breakdown curves, where the composite (SRM) or full (PRM) MS/MS spectra of peptides are acquired while varying the collision energy to capture the intensity of the product ions. In the present account, the pseudo-breakdown curves of 61 stable isotopically labeled (SIL) tryptic peptides corresponding to eight proteins (osteopontin, endoplasmin, glucose-6-phosphatase dehydrogenase, transaldolase, lactate dehydrogenase, alpha actinin 1, filamin A and zyxin), previously identified as non-small cell lung cancer (NSCLC) biomarker candidates (23), were generated by PRM analysis. For "typical" tryptic peptides, i.e., doubly charged peptide comprising 10-16 amino acids, two main scenarios can generally be distinguished. In the first one, illustrated in Figure SI-1A for the peptide EEASDYLELDTIK (m/z 767.374, z = 2+), the abundance of most of the main fragment ions progressively increases with the collision energy to reach a maximum value at nCE 20 (27.63 eV) and then decreases for higher collision energy values. In the second case, more frequently observed, the main fragment ions have various distinct optimum collision energy values. This is illustrated in Figure SI-1B displaying the pseudo-breakdown curves of the peptide AEAGVPAEFSIWTR (m/z 772.393, z = 2+) showing one first optimum collision energy value in the lower range (around nCE 15 (20.85ev)) for two complementary b- and y- type fragment ions generated by facile N-terminal cleavage to a proline residue. These fragments undergo secondary dissociation at higher collision energy while a second optimum value is observed for another set of fragment ions. Although related to a larger number of fragment ions, the second optimum value induces lower overall fragment ion intensities. The evaluation of the impact of the collision energy on the fragmentation pattern of 98 synthetic non-tryptic peptides (including Lys-C, Arg-C, Asp-N and Glu-C peptides) representing the same proteins was performed and allowed similar observation. This is illustrated in Figure SI-1C and D displaying the pseudo-breakdown curves generated for the peptides SILFVPTSAPRGLFDEYGSK (m/z 731.388, z = 3+) and ARVSSGYVPPPVATPFSSK (m/z 489.517, z = 3+) indicating the presence of one or several optimum collision energy values, respectively.

The design of advanced targeted acquisition methods would benefit from optimized fragmentation conditions. In selected reaction monitoring analysis, where each transition can be measured independently using a distinct collision energy value, the optimization of the method is straightforward. For each peptide, the transitions exhibiting the highest intensities are selected and measured using their individual optimum collision energy, which can have a common value for the full selected set or not. By contrast, in parallel reaction monitoring analysis, only a single collision energy value is used to measure each peptide. PRM experiments are generally carried out by applying to the entire set of targeted peptides a unique value of "normalized" collision energy (nCE). A default value of nCE from 25 to 30 has been widely used, as derived from data dependent acquisition (DDA) experiments where it was shown as providing the highest number of peptide identifications by conventional database searching algorithms (29-31). Although it represents a simple approximation, and these values were primarily for identification, i.e. generation of a wide fragmentation pattern, quantitative assays would benefit from fewer more intense fragments. In fact, the MS/MS spectra being associated with a peptide sequence along with the highest Mascot ion scores do not systematically correspond to those where the main fragment ions are measured with the highest intensities, as illustrated in Figure 4. Figure 4A represents the pseudo-breakdown curve of AIPVAQDLNAPSDWDSRGK (m/z 683.348, z = 3+) together with the Mascot ion score in function of the collision energy applied. For this peptide, at low collision energy (CE 17 ev), a few number of multiply charged fragment ions are produced (intense y_{17}^{3+} and in smaller proportions y_{17}^{2+} and y_{15}^{2+}) (Figure 4B). It has been identified with a low peptide Mascot ion score of 22 due to the small number of assigned fragment ions. By increasing the collision energy, the peptide Mascot ion score rises progressively with the increased number of fragments assigned, in spite of lower overall intensity, until a maximum of 86 obtained for a CE of 35 ev (Figure 4C).

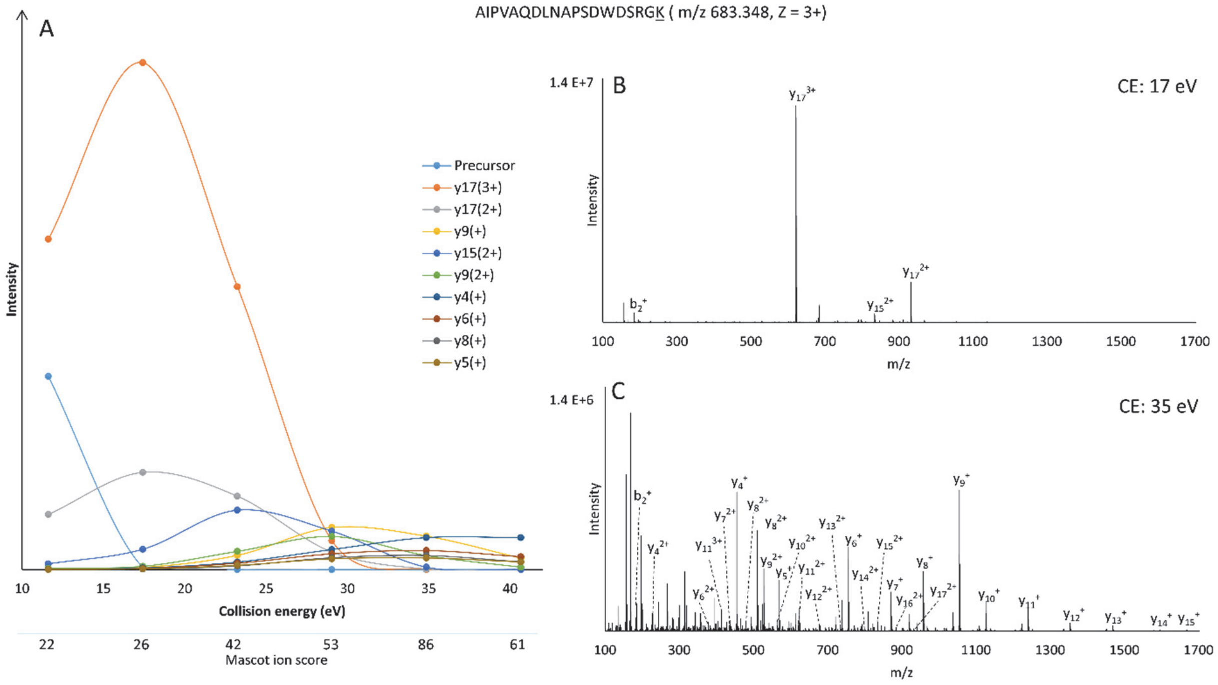


Figure 4: Pseudo break down curve of the AIPVAQDLNAPSDWDSRGK peptide. Fragment ion intensity in function of the collision energy applied and the Mascot ion score for each collision energy (A). MS2 spectrum at CE 17 ev (nCE 10) (B) and at CE 35 ev (nCE 30) (C).

Second, a “normalization” procedure relying only on the mass-to-charge ratio and charge state of the peptides is far too restrictive to really reflect the specificities of the fragmentation process of each peptide. The sensitivity of PRM experiments benefits from a more refined peptide-specific tuning of the collision energy, leveraging the pseudo-breakdown curve information.

In the present account, the determination of the optimum collision energy for PRM analysis of the 159 tryptic and non-tryptic synthetic peptides mentioned *vide supra* was based on their pseudo-breakdown curves. For each peptide the intensity of the most intense fragment ion across the 6 evaluated nCE was compared to that measured at a normalized collision energy of 25 to determine the gain in sensitivity resulting from the fine tuning of the collision energy. The results of this evaluation were grouped by peptide types and are presented in Figure 5.

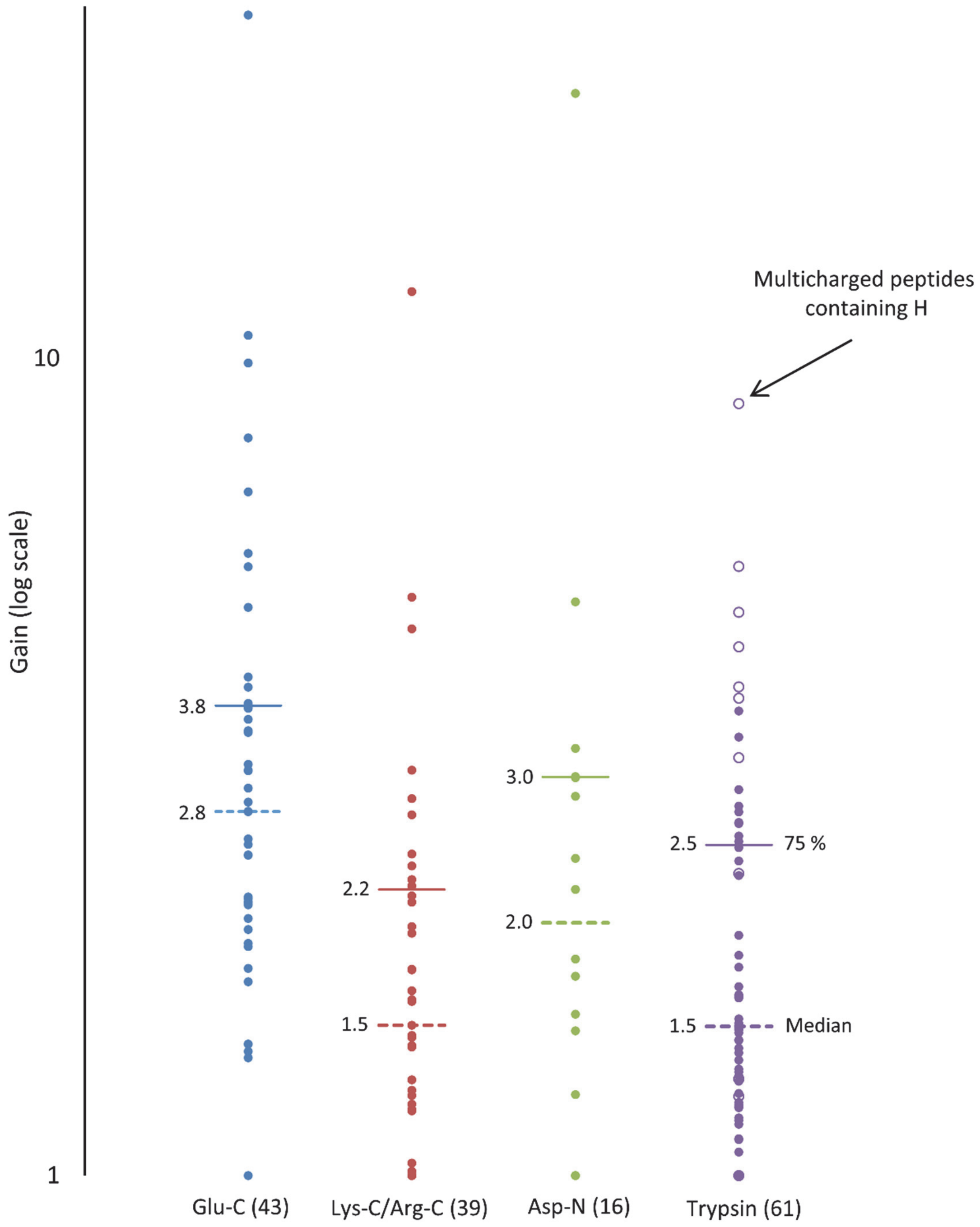


Figure 5: Gain in sensitivity (log scale) of 159 peptides categorized by peptide structure. Based on the breakdown curves of each peptide, the gain was defined as the ratio of the intensity of the most intense fragment ion across the 6 evaluated nCE compared to that measured at the "regular" normalized collision energy of 25 nCE. Dashed lines represent the median value and solid lines represent the upper quartile value.

It turned out that such a peptide-specific optimization of collision energy clearly benefits to the sensitivity of measurement. The gain can be significant (up to 3-10 fold, especially for multiply charged precursors containing additional basic amino acids). All categories combined, for more than a half of the peptides, a minimum gain of sensitivity of two folds was observed.

Conclusion

This study demonstrates that enzymes other than trypsin should be considered when selecting surrogate peptides for targeted proteomics experiments, especially if increased sequence coverage is needed, as they can provide access to information-rich sequences lost during trypsin digestion. This is even more crucial when targeting a certain proteoform for which the PTM or the mutation are not accessible by trypsin. The results show that enzymes cleaving at one site only (such as Lys-C or Arg-C) could have a significant impact on LC-MS density as shown with the digestion of human plasma samples. Finally, the use of a default collision energy, even though allowing higher identification scores during database searches, was shown to have limitations in parallel reaction monitoring-based quantification. The sensitivity of targeted quantification assays can be improved by optimizing the collision energy to produce a few more intense product ions rather than a high number of fragments like it is done for identification.

Associated content

Supplementary information.

Acknowledgement

This work was funded by an AFR (Ref 1194914) grant and a PEARL (CPIL) grant from The Fonds National de la Recherche (FNR). The authors thank the Integrated BioBank of Luxembourg (IBBL), Dr. Guy Berchem (CHL) for providing the plasma samples and Sang-Yoon Kim for technical assistance.

References

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

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

3. Aebersold, R.; Mann, M., Mass spectrometry-based proteomics. *Nature* **2003,** 422, (6928), 198-207.

4. Richards, A. L.; Hebert, A. S.; Ulbrich, A.; Bailey, D. J.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J., One-hour proteome analysis in yeast. *Nat. Protocols* **2015,** 10, (5), 701-714.

5. Beck, S.; Michalski, A.; Raether, O.; Lubeck, M.; Kaspar, S.; Goedecke, N.; Baessmann, C.; Hornburg, D.; Meier, F.; Paron, I.; Kulak, N. A.; Cox, J.; Mann, M., The impact II, a very high resolution quadrupole time-of-flight instrument for deep shotgun proteomics. *Molecular & Cellular Proteomics* **2015**.

6. Tabb, D. L.; Vega-Montoto, L.; Rudnick, P. A.; Variyath, A. M.; Ham, A.-J. L.; Bunk, D. M.; Kilpatrick, L. E.; Billheimer, D. D.; Blackman, R. K.; Cardasis, H. L.; Carr, S. A.; Clauser, K. R.; Jaffe, J. D.; Kowalski, K. A.; Neubert, T. A.; Regnier, F. E.; Schilling, B.; Tegeler, T. J.; Wang, M.; Wang, P.; Whiteaker, J. R.; Zimmerman, L. J.; Fisher, S. J.; Gibson, B. W.; Kinsinger, C. R.; Mesri, M.; Rodriguez, H.; Stein, S. E.; Tempst, P.; Paulovich, A. G.; Liebler, D. C.; Spiegelman, C., Repeatability and Reproducibility in Proteomic Identifications by Liquid Chromatography−Tandem Mass Spectrometry. *Journal of Proteome Research* **2010,** 9, (2), 761-776.

7. Paulovich, A. G.; Billheimer, D.; Ham, A.-J. L.; Vega-Montoto, L.; Rudnick, P. A.; Tabb, D. L.; Wang, P.; Blackman, R. K.; Bunk, D. M.; Cardasis, H. L.; Clauser, K. R.; Kinsinger, C. R.; Schilling, B.; Tegeler, T. J.; Variyath, A. M.; Wang, M.; Whiteaker, J. R.; Zimmerman, L. J.; Fenyo, D.; Carr, S. A.; Fisher, S. J.; Gibson, B. W.; Mesri, M.; Neubert, T. A.; Regnier, F. E.; Rodriguez, H.; Spiegelman, C.; Stein, S. E.; Tempst, P.; Liebler, D. C., Interlaboratory Study Characterizing a Yeast Performance Standard for Benchmarking LC-MS Platform Performance. *Molecular & Cellular Proteomics* **2010,** 9, (2), 242-254.

8. Bell, A. W.; Deutsch, E. W.; Au, C. E.; Kearney, R. E.; Beavis, R.; Sechi, S.; Nilsson, T.; Bergeron, J. J. M., A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat Meth* **2009,** 6, (6), 423-430.

9. Gillet, L. C.; Navarro, P.; Tate, S.; Röst, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R., Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Molecular & Cellular Proteomics* **2012,** 11, (6).

10. Lesur, A.; Domon, B., Advances in high-resolution accurate mass spectrometry application to targeted proteomics. *PROTEOMICS* **2015,** 15, (5-6), 880-890.

11. Silva, J. C.; Gorenstein, M. V.; Li, G. Z.; Vissers, J. P.; Geromanos, S. J., Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics* **2006,** 5, (1), 144-56.

12. Geromanos, S. J.; Vissers, J. P.; Silva, J. C.; Dorschel, C. A.; Li, G. Z.; Gorenstein, M. V.; Bateman, R. H.; Langridge, J. I., The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS. *Proteomics* **2009,** 9, (6), 1683-95.

13. Li, G. Z.; Vissers, J. P.; Silva, J. C.; Golick, D.; Gorenstein, M. V.; Geromanos, S. J., Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures. *Proteomics* **2009,** 9, (6), 1696-719.

14. Gallien, S.; Duriez, E.; Demeure, K.; Domon, B., Selectivity of LC-MS/MS analysis: implication for proteomics experiments. *J Proteomics* **2013,** 81, 148-58.

15. Keller, A.; Bader, S. L.; Shteynberg, D.; Hood, L.; Moritz, R. L., Automated Validation of Results and Removal of Fragment Ion Interferences in Targeted Analysis of Data-independent Acquisition Mass Spectrometry (MS) using SWATHProphet. *Molecular & Cellular Proteomics* **2015,** 14, (5), 1411-1418.

16. Gallien, S.; Duriez, E.; Domon, B., Selected reaction monitoring applied to proteomics. *J Mass Spectrom* **2011,** 46, (3), 298-312.

17. Gallien, S.; Domon, B., Quantitative proteomics using the high resolution accurate mass capabilities of the quadrupole-orbitrap mass spectrometer. *Bioanalysis* **2014,** 6, (16), 2159-70.

18. Gallien, S.; Kim, S. Y.; Domon, B., Large-Scale Targeted Proteomics Using Internal Standard Triggered-Parallel Reaction Monitoring (IS-PRM). *Molecular & Cellular Proteomics* **2015,** 14, (6), 1630-1644.

19. Allen, G.; Campbell, R. O., Specific cleavage of histidine-containing peptides by copper (II). *International Journal of Peptide and Protein Research* **1996,** 48, (3), 265-273.

20. Cedano, J.; Aloy, P.; Perez-Pons, J. A.; Querol, E., Relation between amino acid composition and cellular location of proteins. *J Mol Biol* **1997,** 266, (3), 594-600.

21. Laskay, U. A.; Lobas, A. A.; Srzentic, K.; Gorshkov, M. V.; Tsybin, Y. O., Proteome digestion specificity analysis for rational design of extended bottom-up and middle-down proteomics experiments. *J Proteome Res* **2013,** 12, (12), 5558-69.

22. Guo, X.; Trudgian, D. C.; Lemoff, A.; Yadavalli, S.; Mirzaei, H., Confetti: a multiprotease map of the HeLa proteome for comprehensive proteomics. *Mol Cell Proteomics* **2014,** 13, (6), 1573-84.

23. Kim, Y. J.; Sertamo, K.; Pierrard, M. A.; Mesmin, C.; Kim, S. Y.; Schlesser, M.; Berchem, G.; Domon, B., Verification of the Biomarker Candidates for Non-small-cell Lung Cancer Using a Targeted Proteomics Approach. *J Proteome Res* **2015,** 14, (3), 1412-9.

24. Palagi, P. M.; Walther, D.; Quadroni, M.; Catherinet, S.; Burgess, J.; Zimmermann-Ivol, C. G.; Sanchez, J.-C.; Binz, P.-A.; Hochstrasser, D. F.; Appel, R. D., MSight: An image analysis software for liquid chromatography-mass spectrometry. *PROTEOMICS* **2005,** 5, (9), 2381-2384.

25. Krokhin, O. V.; Spicer, V., Predicting peptide retention times for proteomics. *Curr Protoc Bioinformatics* **2010,** Chapter 13, Unit 13 14.

26. Searle, B. C.; Egertson, J. D.; Bollinger, J. G.; Stergachis, A. B.; MacCoss, M. J., Using data independent acquisition to model high-responding peptides for targeted proteomics experiments. *Molecular & Cellular Proteomics* **2015**.

27. Gupta, N.; Hixson, K. K.; Culley, D. E.; Smith, R. D.; Pevzner, P. A., Analyzing protease specificity and detecting in vivo proteolytic events using tandem mass spectrometry. *Proteomics* **2010,** 10, (15), 2833-44.

28. Lesur, A.; Ancheva, L.; Kim, Y. J.; Berchem, G.; van Oostrum, J.; Domon, B., Screening protein isoforms predictive for cancer using immunoaffinity capture and fast LC-MS in PRM mode. *PROTEOMICS – Clinical Applications* **2015**, n/a-n/a.

29. Kelstrup, C. D.; Jersie-Christensen, R. R.; Batth, T. S.; Arrey, T. N.; Kuehn, A.; Kellmann, M.; Olsen, J. V., Rapid and Deep Proteomes by Faster Sequencing on a Benchtop Quadrupole Ultra-High-Field Orbitrap Mass Spectrometer. *Journal of Proteome Research* **2014,** 13, (12), 6187-6195.

30. Scheltema, R. A.; Hauschild, J.-P.; Lange, O.; Hornburg, D.; Denisov, E.; Damoc, E.; Kuehn, A.; Makarov, A.; Mann, M., The Q Exactive HF, a Benchtop Mass Spectrometer with a Pre-filter, High Performance Quadrupole and an Ultra-High Field Orbitrap Analyzer. *Molecular & Cellular Proteomics* **2014**.

31. Michalski, A.; Damoc, E.; Hauschild, J.-P.; Lange, O.; Wieghaus, A.; Makarov, A.; Nagaraj, N.; Cox, J.; Mann, M.; Horning, S., Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer. *Molecular & Cellular Proteomics* **2011,** 10, (9).