PhD-FSCT-2018-10
The Faculty of Sciences, Technology and Communication

DISSERTATION

Defense held on 16/01/2018 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN CHIMIE

by

Lina ANCHEVA

Born on 5 September 1985 in Skopje, (R Macedonia)

ANALYSIS OF CANCER RELATED PROTEIN ISOFORMS BY MASS SPECTROMETRY

Dissertation defense committee

Dr Serge Haan, dissertation supervisor
Professor, Université du Luxembourg

Dr Dobrin Nedelkov
Isoformix Inc., AZ, USA

Dr Iris Behrmann, Chairman
Professor, Université du Luxembourg

Dr Florian Stengel
Professor, Université du Constans, Germany

Dr Gunnar Dittmar, Vice Chairman
Luxembourg Institute of Health, Luxembour

Affidavit

I hereby confirm that the PhD thesis entitled “Analysis of cancer related protein isoforms by mass spectrometry” has been written independently and without any other sources than cited.

Luxembourg, 15.12.2017 Lina Ancheva

Table of Contents

List of abbreviations:

aa – amino acid

aCGH - array Comparative Genomic Hybridization

AKT – Protein kinase B

ALK – Anaplastic Lymphoma Kinase

CNV – Copy Number Variation

Cq – quantification cycle (PCR)

CV – coefficient of variation

DDM – n-dodecyl-D-maltoside

dHPLC – denaturing High Performance Liquid Chromatography

DMP – dimethyl pimelimidate

DNA – deoxyribonucleic acid

EDTA - ethylenediaminetetraacetic acid

EGFR – epidermal growth factor

ELISA – Enzyme-linked Immunosorbent Assay

ERK – extracellular signal-regulated kinase

ESI – electrospray ionization

FISH – Fluorescent *in situ* Hybridization

GluC – Endopeptidase (serine proteinase)

Grb2 – Growth factor receptor-bound protein 2

HKG – housekeeping gene

HRas – Harvey sarcoma virus oncogene homolog

HVR – hyper variable region

IgG – immunoglobulin

IHC – immunohistochemistry

IP – immunopurification

JAK – Janus kinase

KRas - Kirsten rat sarcoma 2 viral oncogene homolog

LC – liquid chromatography

LCC - Large Cell Carcinoma

LOD – limit of detection

LOQ – limit of quantification

MALDI – Matrix Assisted Laser Desorption/Ionization

MASI – mutant allele-specific imbalance

MSIA – mass spectrometric immunoassay

MEK – Mitogen-activated protein kinase

MET – tyrosine-protein kinase Met or hepatocyte growth factor receptor

mRNA – messenger Ribonucleic Acid

MS – mass spectrometry

mTOR – mechanistic target of rapamycin

nCE – normalized Collision Energy

NGS – Next Generation Sequencing

NOG – n-octyl-β-D-glucoside

NRas - neuroblastoma RAS viral oncogene homolog

NSCLC - non-small cell lung cancer

PCR – Polymerase Chain Reaction

PI3K – Phosphatidylinositol-4,5-bisphopshate 3-kinase

PRM – Parallel Reaction Monitoring

PTM – post-translational modification

RAF – proto-oncogene serine/threonine-protein kinase

RNA – ribonucleic acid

ROS1 – proto-oncogene tyrosine-protein kinase

RTK – Receptor Tyrosine kinase

RT-PCR – Reverse-transcription polymerase chain reaction (or Real-time PCR)

SAA – serum amyloid A

SCC - Squamous Cell Carcinoma

SCLC - Small Cell Lung Cancer

SDS-PAGE – Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis

Shc – Src homology 2-domain

SIL – Stable Isotope Labelled peptides

SISCAPA – Stable Isotope Standards and Capture by Anti-Peptide Antibodies

SNP – Single-Nucleotide Polymorphism

SPE – Solid Phase Extraction

STAT – Signal Transducer and Activator of Transcription

TK – tyrosine kinase

TKI – Tyrosine kinase inhibitor

TM – transmembrane domain

Tyr – tyrosine

WB – Western Blot

SUMMARY

Cancers are frequently caused by protein modifications affecting the biochemical function, resulting in abnormal expression of protein products and resulting in alterations in the cells. Consequently, the study of proteins and their variants constitutes the essence of cancer proteomics aiming at the discovery of novel biomarkers. Changes at the protein level can be introduced mainly by mutations at the DNA level or by post-translational modification events. Although germ-line mutations can result in predisposition to heritable cancers, somatic mutations represent the majority of observed genomic alterations and are not found in matched normal tissues from the same patient. The ability to detect these modified biomolecules with high precision across multiple biological samples is preferentially performed by mass spectrometry based technologies due to the ability of this method to determine the molecular mass of the protein as an accurate characteristic of each protein.

Lung cancer is a heterogeneous disease that is characterized by a spectrum of somatic and genomic “driver” alterations. Profiling lung cancer includes screening for diverse prognostic and predictive biomarkers to assess the prognosis and predict the outcome of treatment. Currently in the clinical environment the treatment options for lung cancer patients are based on histology results and the stage of the tumor. Although these drug treatments demonstrated promising results, patients with an advanced disease stage have poor prognosis also due to an acquired resistance to the drugs. Therefore, EGFR deletion (*del*E746-A750) and point (L858R) mutations (increased or decreased sensitivity to drugs) were used as proof-of-principle for developing a targeted strategy. Furthermore, this strategy was applied on KRas as one of the Ras family isoforms (decreased sensitivity towards EGFR inhibitors) and another “driver” oncogene in NSCLC.

Quantitative mass spectrometric analyses of modified proteins in extracts from tissue samples are however challenging due to the high complexity of the samples and the requirement to detect exactly the modified part of the molecules of interest. Thus, the aim of the project is to design, implement and validate specific tests for protein variants, derived from genetic and/or post-translational changes, known to be involved in cancer formation. The limitations for quantitative mass spectrometry analyses in complex biological samples can be overcome by the use of internal standards in combination with an immuno-enrichment strategy for those proteins containing sequence variations and/or post-

translational modification. Each step was carefully designed to obtain optimal datasets yielding the basis for a solid data analysis of the targeted isoforms. The comparison to the current genomic techniques underlined the importance of the investigation to be performed at the protein level. The creation of a platform with the latest technology advances in mass spectrometry can position these proteomics assays into routine applications and may find an immediate clinical function for patient stratification and ultimately for therapeutic decisions by clinicians.

INTRODUCTION

INTRODUCTION

Somatic mutations and cancer

Cancer as a heterogeneous disease is caused, among others, by DNA alterations that affect the biochemical function or expression of particular genes leading to expansion capabilities to the cell. In the human genome, various types of genes control cell growth in a precise way and once an error occurs in the DNA encoding these genes, their function maybe disturbed and are called “altered” or mutated. The alterations which drive malignancy are characteristic for each cancer and differ between cancer types and even subtypes. These heterogeneous changes can occur as point mutations, insertions, deletions, amplifications, inversions or polymorphisms of the DNA sequence or as epigenetic abnormalities affecting the gene expression [1, 2]. Mutated genes that contribute to malignant transformation can be divided into three main groups: (1) oncogenes - growth promotors, regulators of cell proliferation, apoptosis and differentiation, (2) tumor suppressor genes - inhibitors of cell growth promoting differentiation or stimulating apoptosis and (3) genes involved in DNA repair [3]. The oncogenes as signaling biomolecules involved in signal transduction typically occur in a heterozygous setting [4] and if mutated may become activated leading towards constant cell growth. On the contrary, the tumor suppressor genes being typically homozygous in nature and involved in internal regulatory circuits, and if mutated result in a deletion or inactivation of the genes [5]. Thus, the majority of the targeted anticancer drugs are directed against the activated oncogenes which need to be inhibited. Drugs against tumor suppressor genes, which require an activation instead of an inhibition, are much more difficult to develop and are not really available [6]. Lastly, damages in the DNA repair genes usually occur in hereditary cancers due to various endogenous and/or exogenous factors [7] and can lead to different types of alterations affecting the oncogenes and tumor suppressor genes [8].

When an alteration or a mutation in a gene is present in the germ cells, referred to as a “germline mutation”, it is inheritable and present in all cells [9]. Contrary to the germline variants, the somatic (acquired) mutations in the malignant tissue play a prominent role in the tumor formation and development and these mutations are consequently not found in matching normal tissue from the same patient [10]. An accumulation of somatic mutations in different genes over time is required to cause malignancy. Hence, each cancer is characterized by a set of somatic mutations, of which only a subset contributes to the

tumor's progression. The fraction of alterations responsible for tumor initiation and progression are called "driver" mutations [11], with the remaining part of the mutations called "passenger" mutations, lacking an apparent growth advantage. "Drivers" usually occur in a heterozygous setting in the oncogenes, where one wild-type and one mutant allele are present, making their identification difficult due to the lower expression frequency of the mutation vs. the wild-type expression rate. The solid tumors may contain over a 100 alterations distributed over the coding regions, but only about 5-15 of those are considered to be "drivers" [12]. Distinguishing "driver" mutations from the domination of the neutral "passenger" mutations that characterize each cancer, is important for better understanding of the cancer disease and its subsequent treatment [13, 14]. Furthermore, the "drivers" generally cluster together compared to the "passengers" which are randomly distributed in the genes [15]. These somatic aberrations vary dramatically across cancer types in numbers, patterns and mutation rate [16]. For their analysis, the tissue handling and preservation, the tissue heterogeneity, the approach used for sampling and fixation, the amount of material available for examination and its composition as well as the stage of the cancer (growth, invasion, metastasis) has to be taken into account [17].

As genes are differentially expressed in different cells under various conditions, their products, the proteins, have unique expression patterns too. They are also involved in different signaling pathways, transmitting for example information from the cell surface to the nucleus. Along these signal transductions numerous alterations may occur turning a normal cell into a cancerous one and thus altering its structure and function. Besides the mutations previously mentioned, different post-translational modifications (PTMs) such as phosphorylation, glycosylation, ubiquitination *etc*., are included as well. Therefore, comparison of different cell types (e.g. diseased vs. healthy, treated vs. untreated) for identification of the present mutations and PTMs involved in the signaling pathways and protein-protein interactions is beneficial for the development of targeted therapies, especially when heterozygous mutations are present in the active domains of each gene/protein [18].

Predictive molecular profiling

Molecular profiling represents a method of testing for studying the genetic characteristics of each patient's tumor and searching for unique biomarkers related to that cancer type [19].

The obtained information might be used for identification and development of targeted therapies designed specifically for the analyzed tumor profile. Therefore, based on the patient's profile a corresponding drug type and dosage should be given and screening for additional occurring mutations that may cause resistance to the drug has to be performed. The molecular profiling – including analyses of the tumor's genomic expression and variations, the detection of somatic mutations and/or the activated cellular pathways – could identify patients with specific mutation profiles that could benefit from a specific drug. The acquired diagnostic, prognostic and predictive informations possess clinical significance for improvement of therapies and minimization of "side effects" [20]. For example, patient carrying an activating mutation in the KRas gene (a gene involved in the EGFR's downstream signaling pathways), when treated with an EGFR tyrosine kinase inhibitor or EGFR-directed monoclonal antibody will not benefit from this therapy (figure 1). The unresponsiveness to the given therapy is a result of the activation of a mutation localized "downstream" of the drug target [21].

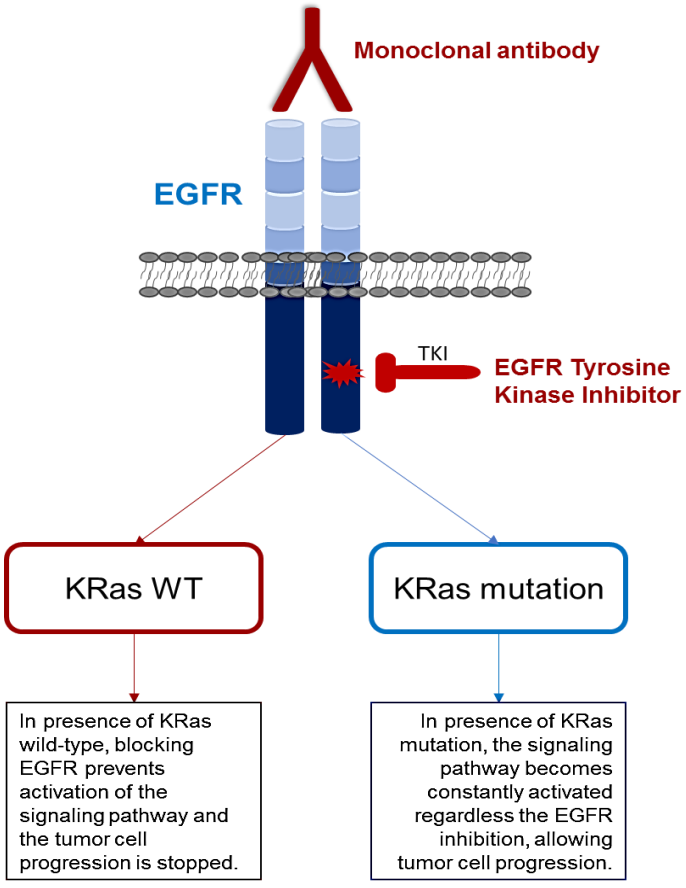


Figure 1. Schematic representation of predicting anti-EGFR treatment efficiency in patients carrying KRas mutation.

For matching the “right” patient with the corresponding drug a complete description of the disease is required as a basis for the patient’s prognosis and/or prediction. In the clinical environment different assays are used for identification of specific altered biomolecules involved in the cell’s regulatory processes, known as biomarkers. These biomarkers are categorized into three main groups: (1) prognostic markers - predictors of the disease outcome independently of the therapy; (2) predictive markers - predictors of the response to a specific therapy; and (3) pharmacodynamic markers - providing answers regarding the therapeutic influence on the patient and if toxic side effects can developed during treatment [22, 23]. By measuring these indicators for a specific biological condition the patients can be easily divided into subgroups, according to their responsiveness, with the creation of a suitable therapy with lower costs and improved clinical benefits. For instance, for non-small cell lung cancer occurrence of mutations in the EGFR protein has a predictive value regarding the responsiveness of the patient to either gefitinib or erlotinib treatment (known EGFR tyrosine kinase inhibitors) [24].

The molecular profiling for specific alterations mainly involves the investigation into genomic aberrations – such as gene fusions, amplifications and/or deletions, copy number variations, gene expression, and DNA methylation as a common epigenetic alteration – at DNA and RNA levels [25]. These traditional examinations search for occurrence of driver mutations in individual genes and monitor their changes and expression in order to predict the tumor’s response to a targeted therapy. Therefore, patients are selected according to the acquired genomic profiles, whereas the prediction of the treatment is based on the drug efficacy targeting the specific protein [26]. These targeted therapies are demonstrating encouraging results, but the overall prognosis remains poor in patients with advanced disease [27] mainly due to the development of secondary mutation resulting in resistance toward the TKIs [28]. Nonetheless, the genomic and transcriptomic events do not always correlate with the protein abundance and expression [29]. This is due to the cellular protein production and maintenance processes such as transcription, processing and degradation of mRNA, translation, localization and post-translational modification events that result in a broad range of protein abundance and expression [30]. Ultimately, proteins are the leading machinery of the cell. With improved analyses techniques, higher sequence coverage is obtained allowing better detection and identification of the protein variations at the amino acid level as well as the post-translational modification changes and their different isoforms. Combining the investigations of the genetic expression, protein profiles, cellular

pathways and existing modifications, an accurate cancer characterization can be made, and diagnostic, prognostic or predictive insights of the disease will be achieved [18], leading to faster diagnosis and accurate treatment choices. Thus, molecular profiling of the protein alterations which have an impact on the disease progression can contribute to the development of individual tailored drugs and avoid the side effects of “one size fits all” treatments.

One limitation during the genomic and proteomic investigation is the specimen itself. Tumor tissue samples are never homogenous [31]. Besides the cancer cells they contain also sections of the normal tissue which reduces the usability of the overall sample amount. This mixture of the normal and cancer cells makes the identification difficult requiring high sensitivity to detect the mutated fraction of the tumor cells [17, 20]. On the other hand, for the molecular profiling, the amount of tissue for analysis is crucial. The necessity to detect and identify all the present alterations within the sample requires the availability of larger sample amounts for different assays. Therefore, cost-effective, multiplexed assays with high-throughput and high sensitivity and selectivity are needed for screening driver mutations at the protein level with a requirement for a minimal amount of sample.

Lung cancer and the main predictive biomarkers

Lung cancer, one of the leading cause of death worldwide, is an example of a heterogeneous disease that is characterized by a spectrum of somatic “driver” alterations. Patients with lung cancer are mainly classified into two groups: non-small cell lung cancer (NSCLC) (accounts for over 80% of all diagnosed lung cancer patients) and small cell lung cancer (SCLC). Further, pathologists divide NSCLC according to the unique molecular signature characteristics in adenocarcinoma (dominant histological subtype), squamous cell carcinoma (SCC) (approx. 33% worldwide) and large cell carcinoma (LCC) (approx. 3% of all lung cancers) [32, 33]. The traditional chemotherapy-based therapies are a first choice of treatment for SCLC, whereas NSCLC patients are less responsive to this regime [34]. If NSCLC is diagnosed at an early stage it can be successfully treated, however, often the disease does not reveal any symptoms for a period of time and it is commonly diagnosed at advanced stage [35]. Currently in the clinical environment different treatment options are offered to lung cancer patients based on histology and tumor stage as well as the patient’s functional ability. Early stage tumors are primarily removed by

surgery and/or treated with platinum-based chemotherapy. On the other hand, due to the variety of somatically acquired mutations, the molecular profiling, diagnosis and management of the advanced NSCLC is done with development of personalized targeted therapies, mainly existent as monoclonal antibodies or as protein kinase inhibitors [36]. Although these drug treatments demonstrated promising results, patients with advanced disease stage have a poor prognosis due to the risk of an acquired resistance to the inhibitors (often caused by development of secondary mutations that cause upregulation of other signaling proteins and pathways) or as a result of disease metastasis [37]. Additional factors that are considered in the tumor profiling and treatment are the patient ethnicity, smoking status, age *etc*. Therefore, for optimal management of targeted therapies for NSCLC, tumors are screened for diverse prognostic and predictive biomarkers to estimate the prognosis and predict the outcome of the treatment.

Profiling lung cancer includes analyses of individual genes for predicting the sensitivity of the tumor towards a drug that targets a specific gene and/or protein. The most involved "driver" genes in NSCLC, as well as colon cancer or breast cancer, are EGFR (increased or decreased sensitivity to drugs), KRas (decreased sensitivity towards EGFR inhibitors), anaplastic lymphoma kinase (ALK) rearrangements, mesenchymal-epithelial transition factor (MET) mutations, phosphatidylinositol-4,5-bisphosphate 3-kinase (PI3K) mutations (increased catalytic efficiency) and proto-oncogene tyrosine-protein kinase (ROS1) rearrangements [38], with a majority having kinase activity and thus becoming desirable targets for anticancer therapies [39]. These targets are involved in a variety of cellular processes like proliferation, motility, suppression of apoptosis and angiogenesis through the RAS/RAF/MEK/ERK, PI3K/AKT/mTOR and JAK/STAT as three major signaling pathways involved in carcinogenesis (figure 2) [34, 40]. Presence of somatic mutations within these pathways turns key components into oncogenes. Patients having specific mutations present within these oncogenes will respond differently to targeted therapies. For example, a patient with an EGFR mutation in the tyrosine kinase domain of the molecule will show a good response to the EGFR tyrosine kinase inhibitors (TKIs), while if a mutation occurs in the KRas gene, the same patient will most probably resist to EGFR inhibitors. For that reason, one of the most affected and investigated alterations in lung cancer signaling pathways is the presence of KRas mutations connected with the smoking or with the occurrence of EGFR mutations related to nonsmoking [41].

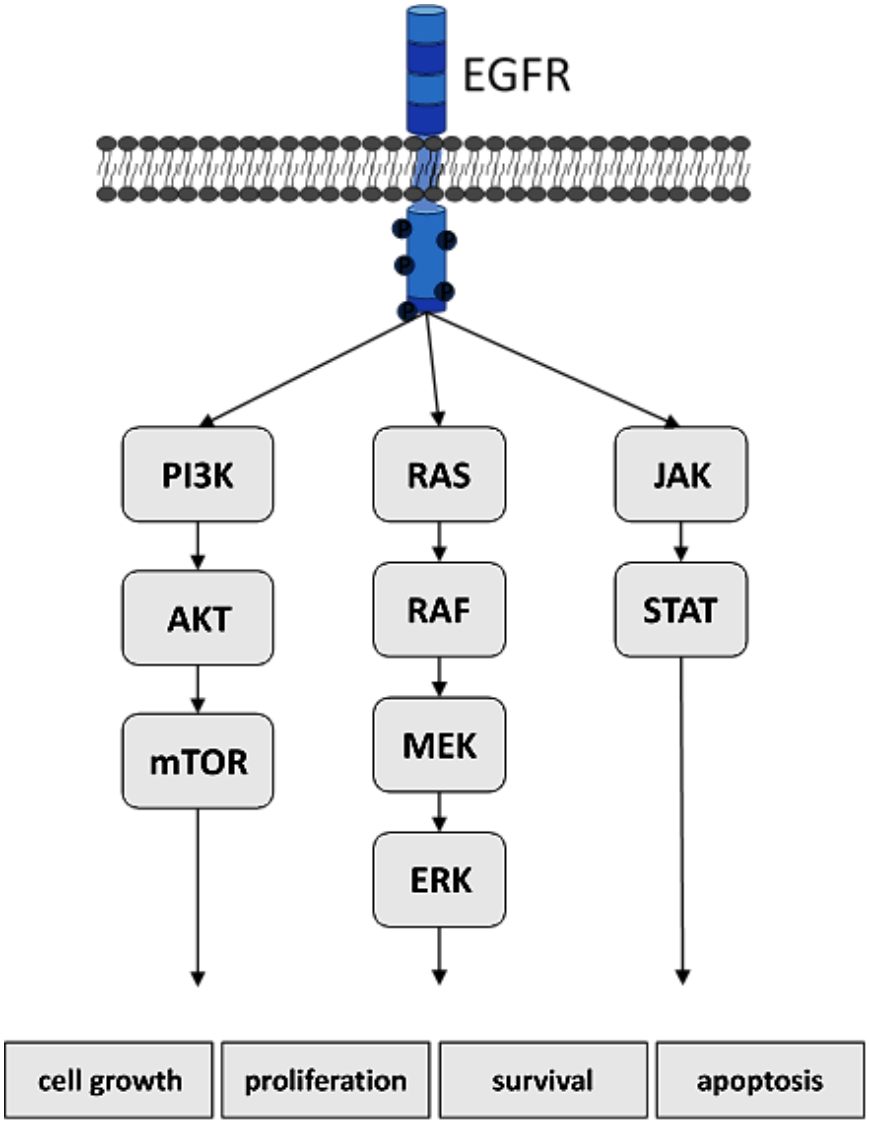


Figure 2. EGFR signaling pathways in lung cancer. Three main downstream signaling pathways involved in cell growth, proliferation, cell survival and apoptosis during signal transduction upon autophosphorylation of key tyrosine sites.

KRas mutations in lung cancer

Kirsten rat sarcoma 2 viral oncogene homolog (KRas) is one of the three Ras family isoform members (the other two are the neuroblastoma RAS viral oncogene homolog (NRas) and the Harvey sarcoma virus (HRas)). It is a small protein with GTPase activity located intracellularly and it is involved in the transduction of the signal from the extracellular through

the intracellular tyrosine kinase domain of EGFR, to the nucleus. KRas is consistent of four domains with 189 amino acids in total. The first domain at the N-terminal part is identical for all RAS isoforms and in this region the most common mutations can occur between amino acids 6 and 16. The second domain is involved in signal transduction and in this region, between amino acids 89 and 97, the three Ras isoforms differentiate in the protein sequence. The first and the second domain together represent the G-domain which has GTPase activity involved in protein-protein interactions. At the C-terminus there is a hypervariable region (HVR), which holds the post-translational modifications involved in membrane anchoring and is responsible for modulating its biological activity (figure 3 and 4).

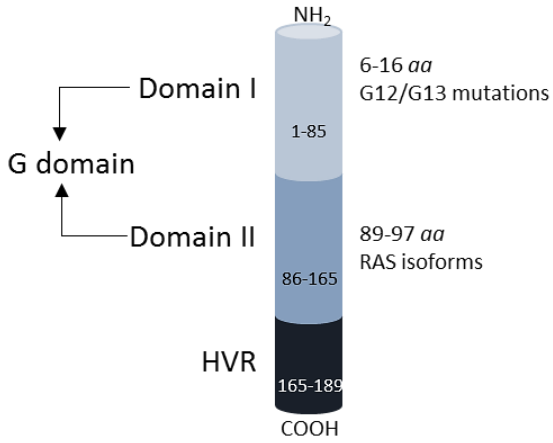


Figure 3. Schematic representation of the RAS structure. Domain I consist of 85 amino acids, domain II of 80, and together they are grouped in G domain. The hyper variable region (HVR) at the C-terminal consists of 24 amino acids.

KRas is a signal transducer involved in cell proliferation, differentiation and survival via different downstream signaling pathways, such as MEK/ERK signaling pathway. KRas is considered to have predictive and prognostic value for various cancer types, like colon, prostate, lung and breast, as well as having an impact on anticancer drug therapies. The KRas (also the other Ras family members) mutational status is important for decision making therapies, especially in NSCLC patients due to the negative impact on the anti-EGFR therapies [42].

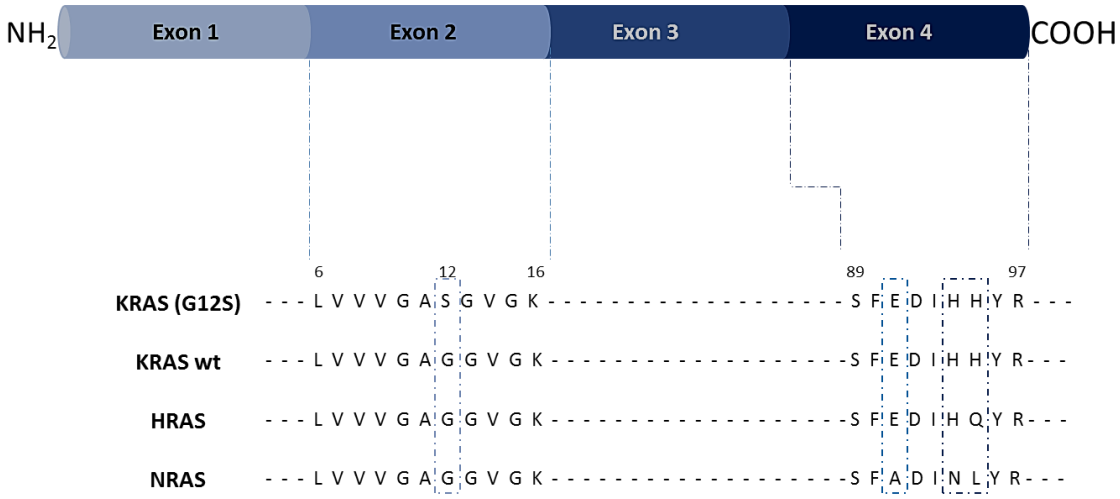


Figure 4. Schematic representation of the KRas structure. In domain I, exon 2, most frequent mutations can occur at codon 12 and 13, between the amino acid 6 and 16 of the KRas protein. Here the sequence for the G12S mutations is presented, characteristic for KRas. The three RAS family isoforms can be distinguished at the C-terminal part of the protein sequence in exon 4, between amino acid 89 and 97 in domain II.

KRas mutations in NSCLC occur mostly at codons 12 and 13, mainly as substitution mutations at position G12A/S/C/D/V or G13D, which reduce the GTPase activity of the protein and thus activate the abovementioned signaling pathways. These heterozygous mutations account for an approximately 30% of the current mutations in lung cancer and are found predominantly in adenocarcinomas and smoker patients with Asian ethnicity [43]. The expression of the KRas, NRas and HRas isoforms varies among species [44] and these highly preserved expression levels have functional importance in the cell proliferation, differentiation and cell death. It is still not clear if the Ras isoform differences are due to the dominant presence of a specific Ras gene in a particular tumor type or as a result of differently translated and expressed Ras protein products displaying diverse biological specificities [44]. For instance, KRas was found as most frequently mutated in colon, pancreatic and lung cancer, whereas mutated NRas is mostly expressed in acute leukemia and the mutation occurrence of HRas was mostly reported in melanoma and bladder carcinomas [45]. Thus, unambiguous determination of the mutation expression of each Ras isoform versus its wild-type counterpart is of functional importance regarding their role in the tumor progression and metastasis.

The prognostic factor of the KRas mutations in NSCLC tumors showed negative values regarding the disease-free and overall survival in surgically treated patients [46]. Moreover, the predictive value of KRas is correlated to the anti-EGFR targeted therapies, although these two mutations (in EGFR and KRas) are mutually exclusive [47]. In other words, the occurrence of KRas mutation in a tumor tissue may indicate absence of EGFR mutations in the tumor. Many studies over the past decades demonstrated KRas as a negative predictive biomarker due to the non-responsiveness of the NSCLC patients carrying KRas mutations when treated with anti-EGFR monoclonal antibodies or inhibitors [48].

EGFR mutations and post-translational modifications in lung cancer

Epidermal growth factor receptor (EGFR) is a receptor tyrosine kinase (RTK) involved in the information transmission from the cell surface to the nucleus. It belongs to the ErbB receptor family and is also known as ERBB1 or HER1. This 170 kDa protein consists of a glycosylated extracellular ligand-binding domain, a transmembrane part and an intracellular domain with tyrosine kinase activity (figure 5) [49]. The extracellular part is divided into four domains, two cysteine-rich and two ligand-binding domains, where the second ligand-binding domain is most involved in dimerization. EGFR is activated by homodimerization with itself or heterodimerization with other HER family members, resulting in structural changes and phosphorylation on the key tyrosine residues associated with increased kinase activity. The tyrosine autophosphorylation allows recruitment of adaptor proteins (e.g. Grb2 or Shc) and these protein-protein interactions trigger downstream intracellular signaling pathways which promote cell growth, proliferation, differentiation, survival and migration [50].

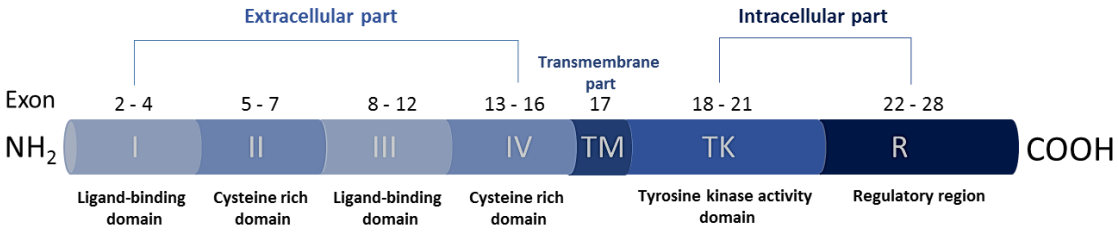


Figure 5. Schematic representation of the EGFR structure. The extracellular part consists of four domains, two ligand-binding (I and III) and two cysteine rich (II and IV) domains. The intracellular part contains a tyrosine kinase domain (TK) and a regulatory (R) region at the C-terminal involved in recruitment of adaptor proteins.

EGFR activation and overexpression can be seen in various tumors, like lung, head and neck, ovary, brain, breast, colon *etc*. [51]. In NSCLC, occurrence of EGFR mutations or protein overexpression were found in about 40-90% of the cases [52]. The activating EGFR mutations are mostly present in the tyrosine kinase domain between exons 18 and 21, mainly as point mutations, deletion mutations and insertions (figure 6). The most common somatic EGFR mutations in lung cancer are found as exon 19 deletions resulting in loss of amino acids 746 to 750 (accounting for 45% of EGFR mutations) and as exon 21 point mutation, resulting in substitution of leucine-to-arginine at amino acid 858 (L858R) (accounting for 40% of EGFR mutations) [53]. These two mutations may result in activation of various cellular signaling pathways – more involved in activation of the antiapoptotic pathways PI3K/AKT and JAK/STAT rather than in ERK/MAPK signaling pathway – leading to cell proliferation or anti-apoptosis [54]. Moreover, these EGFR mutations are considered as predictors of the sensitivity towards EGFR TKIs, thus rendering EGFR as a relevant drug target and defining it as a potential predictive biomarker. The remaining mutations occurring in EGFR can be found as in-frame insertions within exon 20, G719X (X can be C, S or A) point mutations at exon 18, L861Q point mutation at exon 21 and in-frame insertions in exon 19, as well as the T790M mutation in exon 20 that occurs mainly as secondary mutation due to acquired resistance to second-generation anti-EGFR therapies [50, 55].

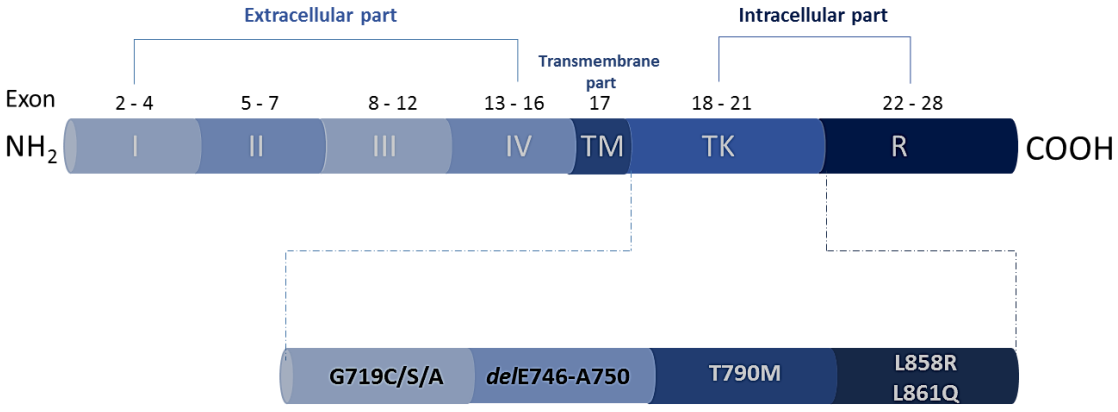


Figure 6. Schematic representation of the main EGFR mutations in NSCLC between exon 18 and 21 in the tyrosine kinase activity domain.

Mutational profiling of EGFR activating alterations is crucial for predicting the sensitivity of targeted therapies. Knowing the levels of the active mutations is important to predict the outcome of the treatment. The receptor can be activated in various ways, including gene mutation, overexpression and/or amplification. The presence of activating mutations in the tyrosine kinase domain of the receptor has a major role in oncogenic determinations in lung cancer and their therapeutic implications [56]. Targeting these mutations with specific drugs helps to improve patient outcomes and overall survival rate. Various targeted strategies have been developed to inhibit EGFR, however gefitinib and erlotinib (two small molecule reversible inhibitor drugs) are the first choice therapeutics that specifically target the EGFR tyrosine kinase activity [57]. The occurrence of activating EGFR mutations in lung adenocarcinomas results in an increased affinity for these inhibitors with an impact on the drug efficiency [58]. Nonetheless, almost all patients develop resistance towards the therapy leading to disease relapse. The acquired resistance is mainly due to the presence of a secondary mutation (e.g. T790M mutation in exon 20), but can also occur following the activation of a parallel downstream signaling pathways, some other phenotypic transformation or as development of a new alteration in addition to the EGFR mutations [59]. Therefore, assessment of the EGFR mutation status and selection of subsets of patient that share characteristic disease profiles (including the tumor stage and type, EGFR amplification and overexpression *etc*.) is important for designing compatible effective therapies.

Other alterations that affect the protein structure and function, causing changes in the amino acid sequence and triggering different cellular signaling processes, impacting the protein localization and interactions are the post-translational modifications. Initiation of phosphorylation as the most frequent PTM in the EGFR protein demonstrated negative influence on anticancer therapeutic targets [60]. The phosphorylation events in the biomolecule are triggered upon activation of the EGFR mutations leading to enhanced phosphorylation onto key phosphorylation sites. Generally, phosphorylation occurs on serine (S), threonine (T) and tyrosine (Y) amino acids (event order S:T:Y = 90:10:0.05). Tyrosine autophosphorylation sites (especially in the regulatory domain of the protein) are the most affected by the activation of the EGFR mutations [61].

EGFR has 20 tyrosine sites prone to autophosphorylation; 10 distributed in the tyrosine kinase domain without major biological significance and 10 localized in the regulatory domain of the protein interacting with various adaptors and signaling proteins (figure 7) [62].

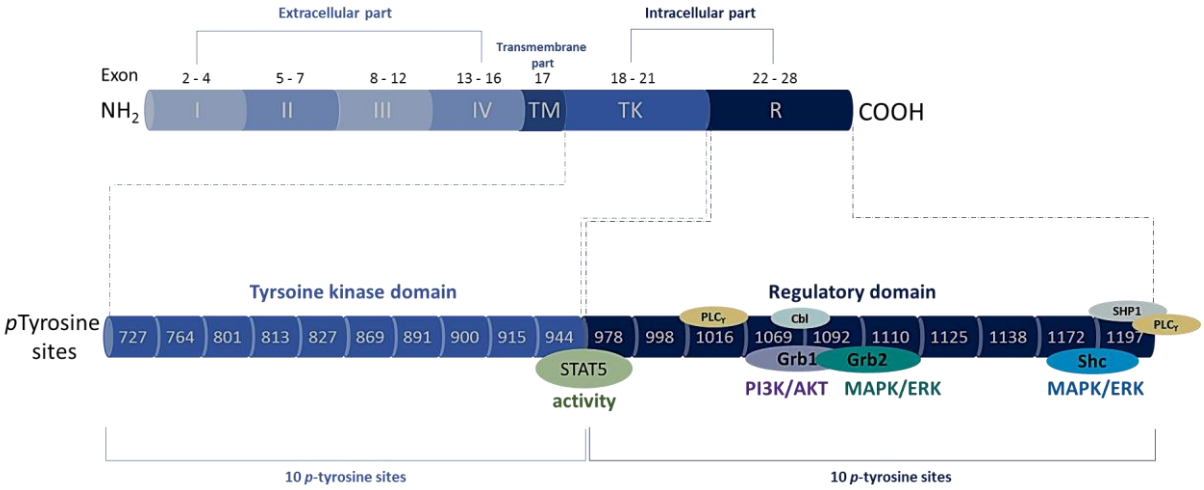


Figure 7. Schematic representation of the 20 phosphotyrosine sites in EGFR protein; 10 located in the tyrosine kinase domain without functional significance and 10 autophosphorylation site located in the regulatory region of EGFR intracellular part.

As mentioned before, phosphorylation and dephosphorylation events on key tyrosine sites can result in recruitment of adaptor proteins of various signaling pathways presenting different signaling outputs of oncogenic processes. Therefore, the ability to precisely identify the activated sites and to determine their kinetic, dynamic and stoichiometry rates as their specific signatures can provide additional insights of the tumor disease and its responsiveness to the TKIs [62].

Mass spectrometry characterization and targeted profiling

Proteins as leading machinery of the cell

Obtaining information for a particular disease includes screening and identification of mutations and their structural variations along with their localization and frequency within the population and their connection to the disease. Detection of the “driver” mutations and

obtaining information for the positive or negative outcomes of targeted therapies is crucial since the rate of the NSCLC patients that can benefit from chemotherapy is very low.

Identification of the important molecular patterns in the tumor cells responsible for the tumor progression and initiation requires precise and reproducible methods that can analyze as many targets as possible at a time. Cancer diagnoses and classification currently are based on the cellular morphology and histological structure. This way information regarding the disease origin, tumor type and stage are obtained as relevant information for patient's therapies. As diagnostic tools various approaches such as genomic sequencing methods, polymerase chain reaction (PCR)-based methodologies, different hybridization techniques and microarrays are used to describe the molecular changes and unique signatures to stratify patients according to their disease occurrence and estimate the survival rate after therapy. These different platforms are using DNAs or RNAs from tumor samples for screening of mutations in the clinical environment. This profiling generally involves estimation of the gene copy number variations (gene amplification or deletions) and the DNA content and expression status. The availability of diverse approaches for investigation of the EGFR mutations, demonstrating different selectivity, resulted in considerable amount of research and results. To date, the estimation of EGFR mutation status by direct sequencing is considered as most clinically relevant method with predictive value [63]. The main limitation of this method is the inability to detect alterations below 25% frequency. Furthermore, PCR-based methodologies alone or combined with next-generation sequencing techniques are broadly used for EGFR genomic studies [64] as well as diverse *in-situ* hybridization methods [65] or immunohistochemistry (IHC) determination of the EGFR overexpression [66]. Although these genomic investigations provided mountains of valuable information, when the results are compared between each-other or with the standard direct sequencing method, inconsistencies might be found such as false-positive or false-negative mutation identification reports [67]. The lack in accuracy across analyses derives from the differences found at the DNA or RNA level with the corresponding protein abundances. They cannot provide information about protein expression and activity, although the proteins are responsible for cellular function. In addition, they also cannot provide any information regarding post-translational modifications such as phosphorylation, glycosylation, methylation, ubiquitination *etc*. Proteins as the drivers of cellular processes make the connection between genomic events and cellular phenotypes. Their expression, shape, charge and function vary between different cells, tissues and microenvironments.

Protein variations involved in different disease processes are expressed as post-translational modifications (PTMs), single nucleotide polymorphisms (SNPs) and alternative splicing forms. Occurrence of these numerous forms makes the proteome much larger than the genome. Besides spliced mRNA transcript variations, SNPs and the PTMs, the different stimulations and environmental factors (*i.e.* temperature, pH, nutrients, cell density *etc*.) make the cellular proteome more dynamic and difficult to predict. This again puts an accent on the importance of the proteins as managers in the regulation of all biological processes. Therefore, proteomic analyses can be useful to obtain all the necessary information regarding significant molecular patterns involved in malignancy and response to therapy, to obtain more accurate information related to disease states as well as identification of novel prognostic or predictive biomarkers with high sensitivity, specificity and reproducibility [68, 69].

Protein abundance and isolation from biological matrices

Quantitative protein assays can describe the protein abundance inside the biological sample and their connection with disease or response to treatment [70]. The biological matrix itself plays a critical point in the analysis of proteins due to its complexity and wide dynamic range. The different behavior of the proteins and peptides can result from the biological variability between the human samples and can have an impact on the overall analyses. Another restraint is the protein abundance in biological samples, like plasma and/or serum (1-4500 mg/dL) [71] or tissue (10^3-10^8 protein copy number/cell) [72]. The mass spectrometric limits of detection (LOD) and quantification (LOQ) of low abundant proteins are in the low pg-to-ng/mL, which makes the detection and quantification of the proteins and peptides difficult. On the other hand, for a mutational profiling and a post-translational modification characterization, larger amount of protein is needed. This is due to the fact that the present alterations represent very small fractions of the total amount of the protein. To overcome these limitations and improve overall analysis of the proteins of interest, especially the low-abundant ones, the preparation of the sample needs to involve a purification step. The traditional proteomics assays use a two dimension-gel (2DG) approach as a separation technique prior MS analysis. Samples even in small volumes can be subjected to electrophoretic separation using dyes or fluorophores as labels for sample visualization, in-gel digestion and sequential mass spectrometric profiling [73]. With this

approach information regarding the protein identification can be obtained together with their molecular weight and quantity as well as a differentiation among mutations and PTMs can be achieved, but with lower sensitivity.

Another way to isolate and identify targeted proteins from the biological matrix is by the immunopurification techniques. This will lead to a decrease in the sample background, improvement of the sensitivity and selectivity of the overall process and avoiding long chromatographic separations and additional fractionations steps. Most of the immunoaffinity-based methods target the proteins of interest by using antibodies to enrich the target(s) from the biological specimens. The antibodies also known as immunoglobulins (Ig), are Y-shaped proteins that interact with a particular protein by forming an antibody-antigen complex. They can be found as polyclonal (having affinity toward same antigen and different epitopes) or monoclonal (having affinity for the same antigen and epitope) antibodies. Antibodies are the entities responsible for the selectivity and sensitivity of immunoassays [74]. The antibodies that recognize the targeted protein can be selective towards the whole protein (known as *pan* antibodies) or against specific peptide sequences with the possible alterations within. This purification step allows enrichment of the targets at protein level or at a peptide level [75]. The purification methodology is based on immobilization of the antibody to a suitable support, capturing the targets followed by a series of mild washes for removing of nonimmunoaffinity-associated components and releasing the protein from the antibody-antigen complex for subsequent analysis [76].

For precise analyses the application of specific antibodies that target particular sequence of the protein can contribute to a better protein identification and characterization. This approach is usually applied for protein post-translational modification analysis, targeting the specific peptides that carry the modification, like ubiquitination [77], phosphorylation [78] or using the general SISCAPA methodology [79]. The latter approach can quantify the protein of interest by using (1) peptide-based antibodies for enrichment of low-abundant peptides as surrogates for the proteins and (2) a synthetic version of the targeted peptide containing stable isotope label in its sequence as control. In contrary, using antibodies against the unmodified part of the sequence, the whole protein can be enriched [80-82]. With this approach higher protein sequence coverage for analysis is obtained, sample complexity is decreased and the wide dynamic range of different protein abundances in the biological matrix such as plasma and tissue is reduced. After this, the analysis can be directed towards

identification and/or quantification of the targeted peptides, identification of the protein mutations or characterization of the post-translational modifications.

Qualitative and quantitative analyses of proteins require high-throughput and great selectivity and sensitivity to be achieved to overcome abovementioned issues. Techniques used to satisfy this requirement have to be very specific towards the target and to be able to precisely distinguish between all the components of the biological matrix. By combining immunoaffinity purification with targeted mass spectrometric analysis greater selectivity and sensitivty are accomplished coming from the antibody and the high-resolution MS-based protein identification and quantification abilities, respectively. Protein isolation from the complex matrix allows further MS determination of the protein itself and the contained amino acid variations and PTMs. Also, reducing the sample complexity prior MS analysis, especially at the beginning of the sample preparation treatment simplifies, facilitates and improves all subsequent steps in the workflow, like digestion efficiency and decreased peptide interferences, better desalting and MS possibility to distinguish between variants with mass differences (*m/z* different values) with low variabilities as well as determination of the abundance of different variant forms from the obtained spectra [83]. Another advantage coming from this merger is that many samples can be treated simultaneously and can be quantified all in one single run.

Proteomic approaches

To cover all the above mentioned parameters, proteomics approaches are divided into two main techniques: (1) Discovery-based proteomic which searches among thousands of proteins within a biological sample to find the unique candidates with prognostic, predictive or therapeutic values, and (2) targeted proteomic which qualitatively and quantitatively measures known protein targets in biological samples in correlation with their cellular functions and protein network interactions. In these approaches, methods as gel electrophoresis, affinity-based techniques, fluorescence techniques and mass spectrometry are applied for detection of the frequency and abundance of different proteins and their isoforms, mainly as post-translational modifications. Majority of these methodologies have demanding sample preparation and processing with low-throughput and inability to identify protein structural modifications [84]. On contrary, mass spectrometry-based methods provide information about the protein mass, its expression, function and structural changes

with high sensitivity, selectivity and high-throughput performances in relation to the genomics and transcriptomics data [85]. This collection of information can be achieved by top-down or bottom-up approaches. The former analyzes intact proteins without previous fragmentation giving results for the whole protein sequence including PTMs identification and site occupancy. The latter method digests proteins into peptides using proteolytic enzymes and investigates these fragments that hold information regarding the protein and its possible variations usually by tandem mass spectrometry [86, 87]. Regarding the two strategies, different labelling methods can be combined to achieve better protein sequence coverage as well as complete PTM identification.

Mass spectrometry-based analysis

As mentioned above, mass spectrometry based proteomic analysis can obtain information regarding protein mutations and post-translational modifications related to disease in a qualitative and quantitative manner by measuring the mass-to-charge (*m/z*) ratio. The basic sample preparation is identical to various bottom-up MS-based approaches and as starting material uses proteins extracted from the biological sample (step 1 in figure 8). The proteins either in solution or in gel are digested by suitable proteolytic enzyme(s) resulting in mixture of representative peptides (step 2). These signature peptide sequences cover the alteration representative for the protein of interest and are further dissolved in a solution that represents the mobile phase for the liquid chromatography system. The LC separation (step 3) is based on interaction between the mobile and stationary phase, where each peptide travels with different speed and elutes from the chromatographic column at different retention time. Next, the eluted peptides enter the ionization source of the mass spectrometer (step 4) where they undergo electro-spray (ESI) or matrix-assisted laser desorption (MALDI) ionization. At this step, multiple charged gaseous ions are produced suitable for resolution and separation in the mass analyzer. In the mass analyzer (step 5) ions are separated according their *m/z* ratio by an electric or magnetic fields, where lighter ions are detected faster compared to the heavier ones. At the end (step 6), the registration and detection of each ion by the detector results in mass spectrum of the fragment ions (Figure 8).

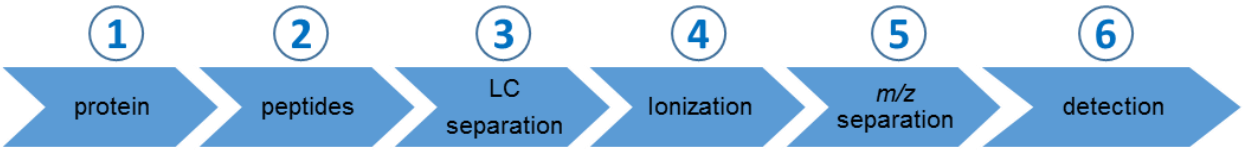


Figure 8. Schematic representation of the complete mass spectrometric analysis (bottom-up). The extracted proteins are digested by suitable protease(s) to generate signature peptides. The representative peptides are then chromatographically separated prior MS analysis. The MS analysis consists of precursor ion selection, its fragmentation and selection of most intense fragment ions, being detected at the end of the analysis.

Targeted proteomic analysis

To identify the targeted protein, its amino acid sequence, mass-to-charge ratio, retention time as well as the intensity and distribution of the fragment ions, high sensitivity and selectivity are required for the chosen methodology. Lately, parallel reaction monitoring or PRM has been successfully applied to discovery and targeted MS-based experiments, providing high specificity and high mass accuracy using small or limited quantities of biological material. With this targeted approach not only the information regarding the targeted protein and its mutational variations can be obtained, but also the post-translationally modified isoforms can be differentiated [88].

The PRM method can be performed on a hybrid quadrupole time-of-flight or quadrupole-Orbitrap mass spectrometer, achieving high specificity and selectivity in high resolution and obtaining full MS/MS spectra containing all product ions of the targeted peptides [89]. The main advantage of PRM is that all the precursor and product ion fragment pairs (also known as transitions) are monitored at the same time (full MS/MS spectra acquiring), increasing the identification of a targeted peptide. Due to this feature, no special prior optimization is necessary as well as no prior knowledge and preselection of the peptides is required. The only information about the precursor *m/z* and the expected elution time are necessary prior analysis regarding the targets. Concerning the instrument, definition of the isolation window, the maximum fill time, the monitoring window and the resolution of the Orbitrap are needed for targeted analysis [88]. Thanks to the flexibility of the instrument, the discovery and targeted analysis can be held at the same time resulting in selection of those targets with highest sensitivity under various conditions and successful detection and quantification of the targeted proteins.

Briefly, in PRM, first precursor ions are isolated during their chromatographic elution on a defined isolation window and afterwards are transferred to the collision cell where they undergo fragmentation at the optimal collision energy characteristic for each peptide, or a normalized one for all peptides. Then, the obtained fragment ions are sent to the C-trap from where they are transferred to the Orbitrap where mass analysis occurs (figure 9) [88].

Here, full MS/MS spectra are acquired for each precursor ion fragments for which all information are available at any time. The acquisition interpretation of the data provides information about the peptide identification and the protein quantification using the most representative transitions with highest sensitivity and selectivity. During data processing, the obtained signals are compared to the signals of the synthetically labelled peptides which serve as internal standards and the identification of the target is confirmed along with its quantification. But, even though everything sounds ideal there are some drawbacks that make the analysis complex [90]. First, the targeted peptides have diverse physicochemical properties (like size, hydrophobicity, and charge state) making them behave differently during ionization and in the mass spectrometer. Second, the biological sample matrix where the peptides are located have an impact on the chromatographic separations of the targets, their ionization and later on the calculated recovery. Third, the sample preparation steps, as enrichment selectivity, digestion efficiency and internal standard spiking contribute to the end result. Nonetheless, having the power to deliver information about the proteins of interest and their structural modifications without any prior knowledge demonstrate mass spectrometry as the most sensitive choice for protein analysis.

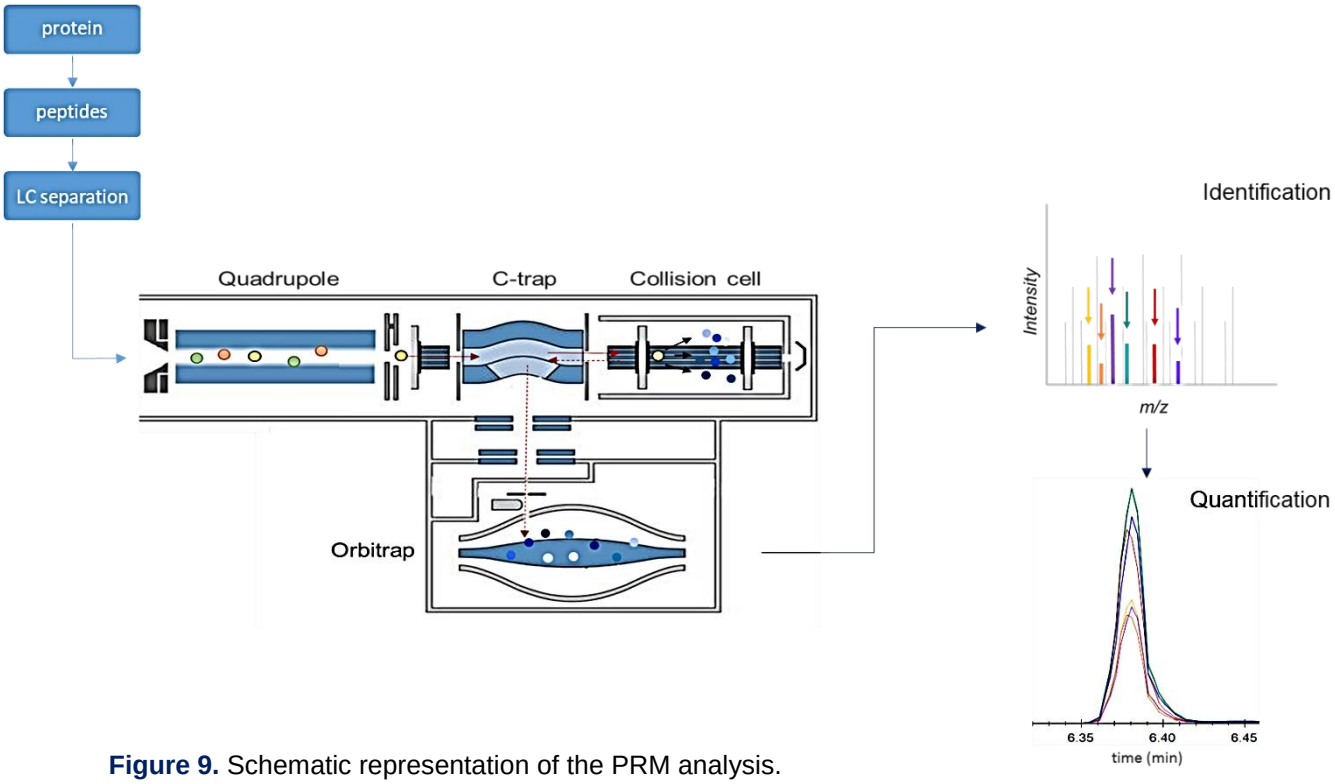


Figure 9. Schematic representation of the PRM analysis.

THESIS OUTLINE

The aim of this project is to design an analytical platform based on immunoaffinity enrichment of the proteins of interest followed by targeted mass spectrometry analysis. After obtaining the enriched proteins from a complex biological matrix the analysis can be directed towards their detection/identification, screening for mutations and/or analyzing the post-translational modifications. The EGFR and KRas proteins have been selected to serve as proof-of-principle as they harbor different types of mutations and post-translational modifications, such as phosphorylation. The proteins were purified from different cancer cell lines, using monoclonal antibodies immobilized on protein A/G micro-columns. After proteolysis, stable isotope labeled peptides were used as internal standards for LC-MS identification and quantification of the peptides.

As described in **Chapter I,** the immunoaffinity method was developed for EGFR and KRas purification from four lung adenocarcinoma cell lines (A549, H1975, HCC827 and H3255), using different antibodies and affinity supports. The method optimization included control of the sensitivity, accuracy, precision and linearity range as critical parameters, monitoring the total EGFR protein expression within each cell line. Three control peptides representing the total EGFR were used as internal standard controls for the PRM analyses. Validation of the designed workflow was achieved by comparison with the membrane fractionation technique as a widely used method for membrane protein isolation.

The immunoaffinity reduces dramatically the complexity of the sample and improves the overall selectivity and sensitivity of the subsequent LC-MS analysis. It allows usage of shorter chromatographic separations (37-minute gradient times) compared to the standard 90 to 180 minute gradients. The developed strategy demonstrated good reproducibility and overall sensitivity and selectivity, with minor limitation with respect to protein recovery. Moreover, the PRM analysis improved the identification and quantification of the targeted peptides in a high-throughput and sensitive manner.

Profiling EGFR and KRas diverse mutations within the same cell line and making a comparison between the different alterations can serve as a model for targeted therapy treatments. The published results (Lesur A., Ancheva L. et al., *PCA*, 2015, given in annex) using cancer cells and tissue, as well as cancer serum plasma samples demonstrated

unambiguous identification of the EGFR, KRas and SAA protein isoforms with a 7-minute separation time prior to PRM analysis. Further, the optimized approach was applied on five different cancer cell lines, carrying different EGFR and KRas mutations as the main study for **Chapter II**. The G12S mutation was identified and quantified, and the differentiation between the three Ras family isoforms expression was performed in the different cell lines. Regarding the EGFR protein, the expression of the deletion and point mutation versus the wild-type counterparts was measured, demonstrating the valuable potential of this approach for providing fast results essential for patient stratification in targeted therapies.

In current clinical settings the identification of the small subset of variations that play an important role in tumor initiations and progression is mainly performed at the genomic level. The copy number variations, DNA content and expression are followed as clinically relevant parameters using the standard sequencing technology or PCR, hybridization-based and/or antibody-based methods. However, the genomic and transcriptomic events do not correlate with the protein abundance and expression (as a drug targets). Therefore, to improve the patient selection and therapy outcomes, investigation of the genomic along with the proteomic profiles is beneficial for accurate prognostic and predictive insights into the disease. Accordingly, in **Chapter III**, the comparison of the EGFR copy number variations, DNA content, mRNA and protein expression, as parameters with predictive value, demonstrated the significance of performing the analysis at protein level due to deeper sequence coverage and obtaining the “real” picture for a disease (manuscript in preparation). The results pointed out that genomic analyses can serve as indicators for identification of potential candidates for sequential protein profiling. Moreover, with the ability to deliver information about the proteins of interest and their structural modifications (such as “driver” mutations and post-translational modifications) without any prior knowledge, mass spectrometry working in PRM mode demonstrated itself as the most sensitive choice for protein analysis.

Since EGFR was selected as a main target of investigation in **Chapter IV** characterization of its tyrosine phosphorylation is presented, to obtain an as complete picture for this protein state as possible. Activation of the EGFR mutations in the tyrosine kinase domain of the

biomolecule enhances the autophosphorylation onto key tyrosine sites as signal transducers. EGFR was purified at the protein level using monoclonal antibody immobilized onto protein A/G affinity support, resulting in a broader sequence coverage, better preservation of the phosphorylated sites and excluding peptide-based enrichment steps. This approach allowed identification of 11 phosphotyrosine sites (out of 20) in four lung adenocarcinoma cells and one epidermoid cancer cell line. Furthermore, estimation of their stoichiometry was achieved to determine the frequency of the protein phosphorylation on each phosphotyrosine site by implementing dephosphorylation step using the alkaline phosphatase enzyme. Six autophosphorylation sites were identified and quantified among all the cells, with highest expression in the two cell lines carrying EGFR point mutations. These analyses demonstrated the strong relationship between the occurrence of the L858R point mutation and activation of the key tyrosine autophosphorylation sites. Additionally, the dynamic profiles were assessed upon EGF stimulation, confirming the increased level of phosphorylation upon stimulation.

In conclusion, an analytical platform for unambiguous characterization of oncoprotein related mutations, post-translational modifications and isoforms was established, combining protein immunopurification with targeted MS analysis. The developed and optimized methodology successfully detected and quantified targeted KRas and EGFR isoforms, frequently descriptive for lung cancer. The application was complemented by the EGFR phosphorylation analysis for their impact on targeted therapies. This high-throughput approach holds clinical relevance for improved patient selection for therapeutic strategies.

Chapter I

Method development and optimization

BACKGROUND

Identifying and quantifying clinically relevant proteins in biomedical investigations requires the usage of specific platforms with high selectivity and sensitivity. Mass spectrometry-based methods with their high-throughput capabilities can be used in different biomarker verification studies [91, 92], and qualitative and quantitative analysis of cellular processes and states [93, 94]. Also, due to their ability to measure multiple analytes in different abundance, volumes and conditions, they can be very useful in the clinical diagnostic environment.

In this chapter, we present a multiplexed strategy for identification and quantification of proteins known to harbor driver mutations. The methodology described below was developed by coupling immunoaffinity purification – method used for the isolation of proteins of interest from complex matrices – with a subsequent targeted mass spectrometry-based analysis. One advantage of this IP-MS merger is that only one antibody is necessary per protein enrichment, omitting additional fractionation steps and simplifying the subsequent analysis like decreasing the chromatographic gradient separation time. The isolation of the protein of interest reduces the ion suppression and interferences from the complex biological matrix and as a result, the differentiation between the isoforms and specific mutations is improved [95] as well as the overall sensitivity and selectivity [96]. Additionally, protein purification can enable the characterization of existing post-translational modifications involved in protein interactions and functions [97].

RESULTS

Method development

The aim of the overall strategy is to purify the proteins of interest from the selected biological samples, identify the driver mutations, estimate their expression level in the cells versus the wild-type and/or to characterize the PTMs in these protein enriched samples by targeted LC-MS analysis.

The developed workflow presented in figure I.1, comprises six independent steps. Each step was individually controlled and optimized using epidermal growth factor receptor (EGFR) as proof-of-principle. Briefly, a biorepository of cancer cell lines harboring different EGFR mutations was established (table I.1). The cells were lysed using a

detergent-based lysis buffer and the proteins were extracted. After enrichment of EGFR with monoclonal antibodies, previously selected representative peptides covering the mutation and wild-type sequences were generated by different proteases. The peptides were then analyzed by a targeted LC-MS method working in parallel reaction monitoring (PRM) mode, using synthetically labeled peptides (SIL) as internal standard controls. Further down, the workflow's step-by-step design and critical parameters are described.

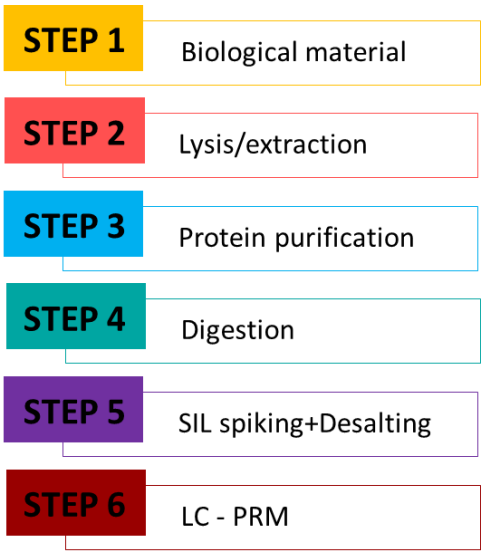


Figure I.1. Developed workflow for immuno-MS targeted analysis using EGFR from lung adenocarcinoma cell lines.

Biological material

Four lung adenocarcinoma cell lines harboring EGFR exon 19 deletion (*del*E746-A750), exon 21-point mutation (L858R) and wild-type (EGFRwt) and one epidermoid cancer cell line overexpressing wild-type EGFR were selected for this study; their characteristics are presented in table I.1.

Cells were harvested in suitable media until they were confluent and collected for subsequent use or frozen as cell pellets for later LC-MS analysis.

Cell lysis and protein extraction

To obtain the proteins of interest from the biological material, cells have to be disrupted to make all the cellular compartments and sub-cellular material accessible for later analysis. The cell lysis step has two critical parameters, the lysis buffer composition and the lysis method, which had to be controlled and optimized. The composition of the lysis buffer depends on the protein localization in the cell. For optimal protein extraction, especially EGFR as a transmembrane protein, the composition of the lysis buffer must include detergents (to break-up the membrane structures), buffers (for pH stabilization), salts (to regulate the acidity and osmolarity of lysate), protease and phosphatase inhibitors (to

preserve protein integrity and function) and other components like glycerol (to preserve the protein folding and possible interactions), reducing agents like dithiothreitol (for protein denaturation) and/or chelating agents like EDTA (to prevent metal ion binding). Regarding the cell lysis, various methods can be applied – including freeze/thaw cycles, mechanical or chemical disruptions as homogenization or sonication and needle passages – for efficient solubilization and isolation of the targeted proteins.

Table I.1. Description of the selected cell lines for this study*.

Cell line	Cancer type	Zygosity	EGFR gene	EGFR mutation type
A549	Lung adenocarcinoma	Homozygous	/	wild-type
H1975	Lung adenocarcinoma	Heterozygous	/	L858R
HCC827	Lung adenocarcinoma	Heterozygous	amplification	overexpression *del*746-750
H3255	Lung adenocarcinoma	Homozygous	amplification	overexpression L858R
A431	Epithelial carcinoma	Homozygous	amplification	overexpression wild-type

*Information obtained from http://cancer.sanger.ac.uk/cell_lines (Catalogue of Somatic Mutations in Cancer Database v82).

For the isolation of the membrane proteins, such as EGFR in our study, detergents, buffers, salts and inhibitors were used to detach the protein from the lipid layer of the cell membrane; protease and phosphatase inhibitors are also necessary to preserve the protein from degradation and modifications caused by these enzymes (e.g. dephosphorylation) [98]. Moreover, the ionic strength of the detergents, their compatibility with the subsequent immunopurification and MS analysis and the pH of the lysis buffer have to be taken into account for optimal cell solubilization and protein extraction. Following these criteria, three lysis buffers, containing 4% octyl-β-D-glucopyranoside (NOG), 1% n-dodecyl β-D-maltoside (DDM) or 1% Digitonin detergents, were compared by western blot (WB) analysis targeting the total EGFR protein in the five cancer cell lines previously described. Recombinant EGFR protein (95 kDa, external part of the protein) was used as a control. As it can be observed

in the gel presented in figure I.2, the DDM lysis buffer extracted the highest amount of EGFR from all the cells and was selected for subsequent analyses.

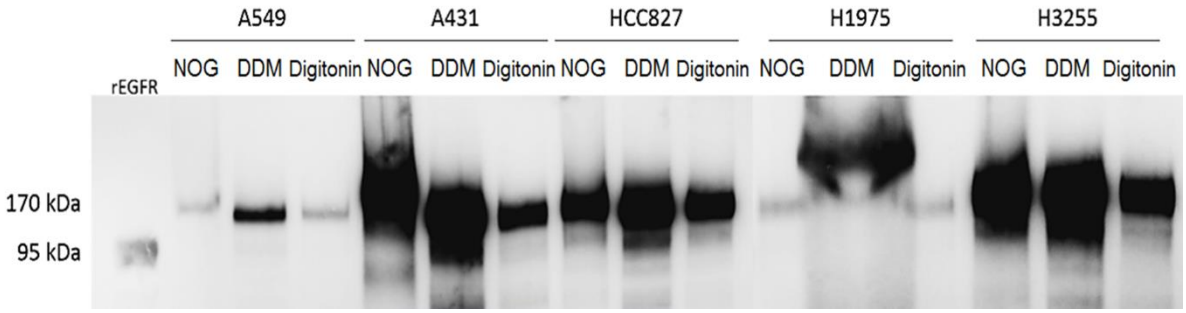


Figure I.2. Comparison of lysis buffers containing 4%NOG, 1%DDM or 1%Digitonin for EGFR extraction. WB analysis of total protein extracts (20 µg) from five different cancer cell lines; 0.5 µg of rEGFR was used as control, targeting the total EGFR from the cells. DDM lysis buffer (middle band) was chosen for subsequent analyses.

Then, three freeze/thaw (-80°/25°C) and five-to-ten homogenization cycles were used as lysis methods in multiple experiments. Both methods lysed the cells efficiently, however the extraction of EGFR was better by homogenizing the cells, a fact observed during multiple IP-MS analysis (figure I.3).

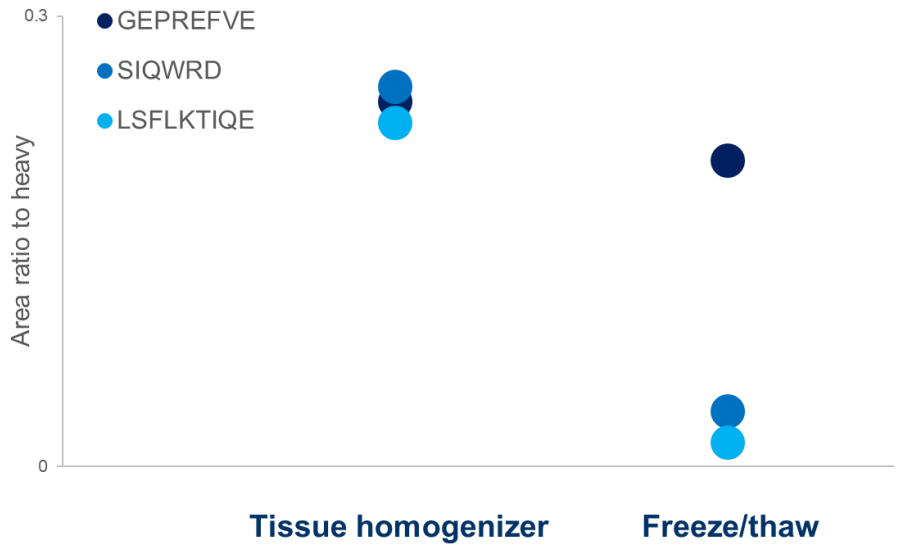


Figure I.3. Comparison of homogenization (left) and freeze-thaw (right) cycles for cell lysis and EGFR extraction. Dots present endogenous/heavy peptide area ratio intensities of the three EGFR control peptides presented as mean values calculated from n=2 replicates from the H3255 cell line. Higher and more reproducible signal intensities were obtained with the homogenization lysis method.

The efficacy of the selected lysis conditions and parameters was evaluated by comparing the WB results with an MS analysis performed on the selected cell extracts submitted to WB analysis. The intensities of the three control peptides representing the total EGFR protein expression in the cells (described in the *Synthetically labelled peptides* section from this chapter) were monitored by LC-MS working in data-dependent acquisition mode (DDA), using SIL peptides as internal standard controls. Measured signals, SIL (upper chromatograms) and endogenous (bottom chromatograms) peptides for EGFR in the H3255 cell line presented as an example in figure I.4, confirmed the protein's presence in the cell extract and consequently the successful cell lysis and protein extraction step.

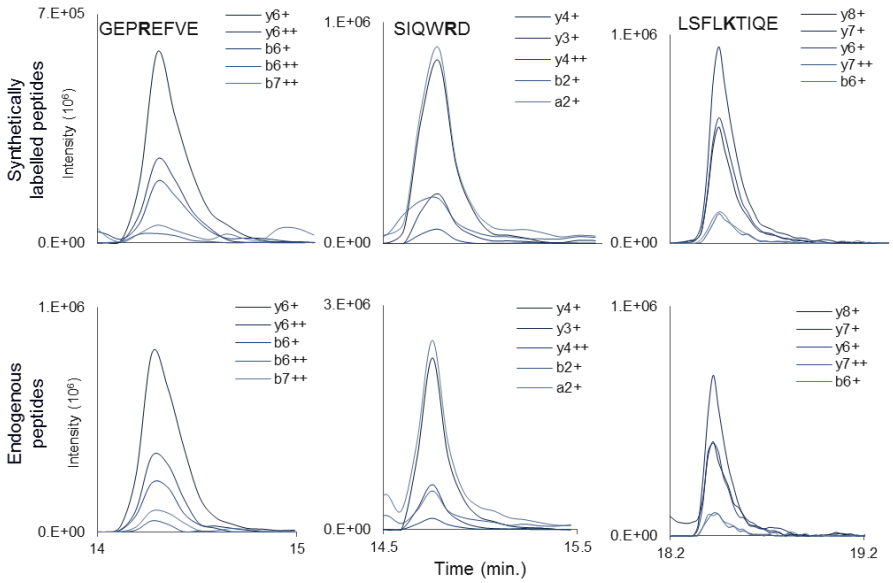


Figure I.4. Chromatograms of the three EGFR control peptides; SIL (upper) and endogenous (bottom); total protein extract (20 µg) from the H3255 cell line was analyzed in DDA mode. Obtained signals demonstrated the presence of EGFR in the cells. The bold arginine (R) and lysine (K) letters indicate the synthetically labelled amino acid.

Protein purification

Sample complexity is one of the main limitations in biochemical analysis due to the negative impact over the detection of targeted alterations mainly coming from the wide dynamic range of the protein abundance. Therefore, purification and enrichment of the proteins of

interest is advantageous resulting in decreased sample complexity and increased material for analysis. The basic immunopurification (IP) workflow includes activation and equilibration of the affinity support, antibody binding onto the support, formation of an antibody-antigen complex, removal of unbounded contaminants and elution of the antigen by breaking the antibody-antigen complex [9]. The choice of an antibody and of a suitable affinity support are the main critical parameters of this step as well as the pH, polarity and ionic strength of the reagents used.

Disposable automated research tips filled with porous support on the entrance of the tip (micro-columns further in the text) were chosen as affinity support, due to the repeated aspiration/dispense cycles allowing closer antibody-antigen contact [100]. Monoclonal anti-EGFR (clone 528) was chosen for EGFR enrichment due to the greater purity and concentration compared to polyclonal antibodies. Various suppliers of this antibody clone – namely Millipore, Thermo and Novus Biologicals – were compared considering the antibody IgG isotype (IgG_2) binding affinity towards the affinity support. Different amounts were tested and two micrograms of antibody were chosen as the adequate amount for binding to the affinity support and protein capture. The highest signals were obtained with the Novus antibody, when the signal intensities of the three control peptides representative of EGFR expression in the cell extracts after IP-PRM analysis were compared (figure I.5).

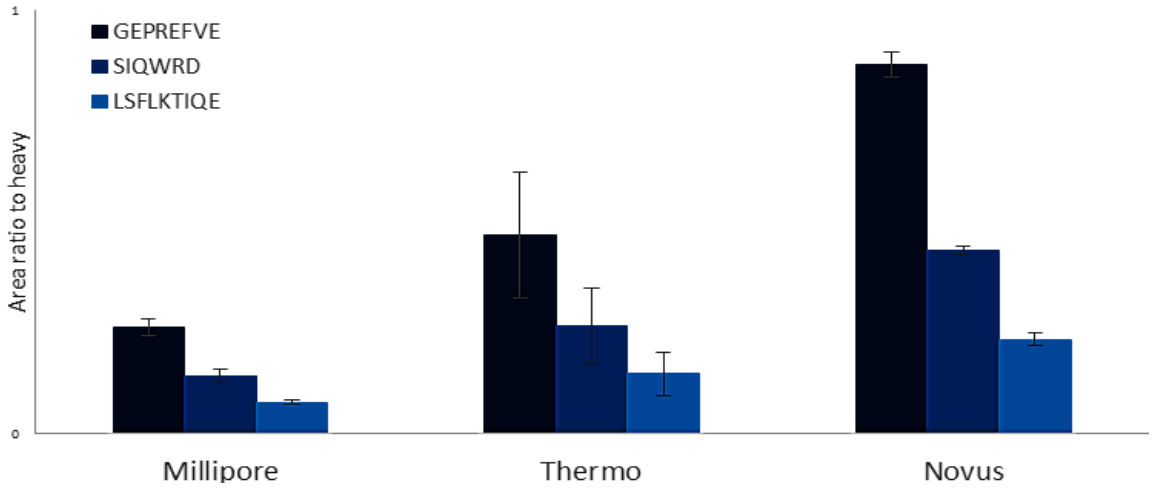


Figure I.5. Comparison of three monoclonal anti-EGFR (clone 528) antibody suppliers. Bars present endogenous/heavy peptide area ratio intensities of the three EGFR control peptides presented as mean ± SD variations calculated from n=2 replicates from the HCC827 cell line. Highest signal intensities were obtained with the antibody supplied by Novus.

After choosing the most suitable antibody supplier, different affinity supports in a micro-column format (porous support placed at the entrance of a tip), composed of protein A (containing four binding sites), protein G (two binding sites) and protein A/G (six binding sites), were initially compared. The selected antibody showed strong binding affinity to all three supports. However, the PRM analysis of the purified EGFR from the four lung adenocarcinoma cell lines showed most optimal binding affinity of the antibody towards the protein A/G micro-column (figure I.6), showing highest signals for the EGFR control peptides in majority of the samples.

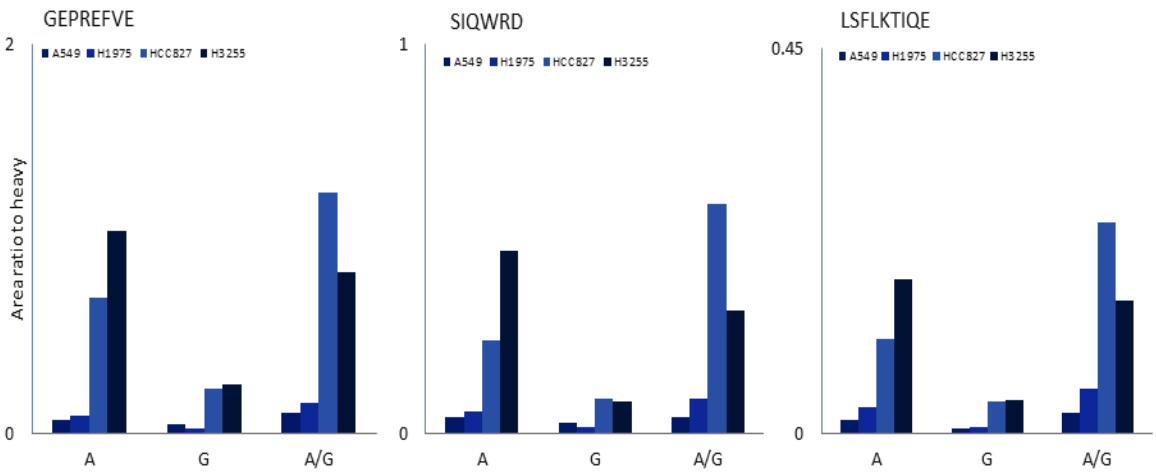


Figure I.6. Comparison of protein A, protein G and protein A/G affinity supports using anti-EGFR (clone 528) antibody supplied by Novus. Bars present endogenous/heavy peptide area ratio intensities of the three EGFR control peptides presented as mean values calculated from n=2 replicates from four lung adenocarcinoma cell lines. Protein A/G micro-columns displayed highest signal intensities in all cell lines.

To confirm the protein A/G affinity performance, this affinity support was additionally compared to protein G magnetic beads and Streptavidin magnetic beads and micro-columns. The Novus antibody was used for protein A/G and protein G supports, whereas a biotinylated anti-EGFR (clone 528) antibody immobilized onto Streptavidin supports, was used towards EGFR from A549 and H1975 cell lines. Although higher signals were expected from the biotin-streptavidin ligand-binding interaction due to their greater affinity, the observed variabilities between the replicates were much higher than with the protein

A/G support (figure I.7), confirming the choice of protein A/G as the most suitable affinity support for EGFR purification.

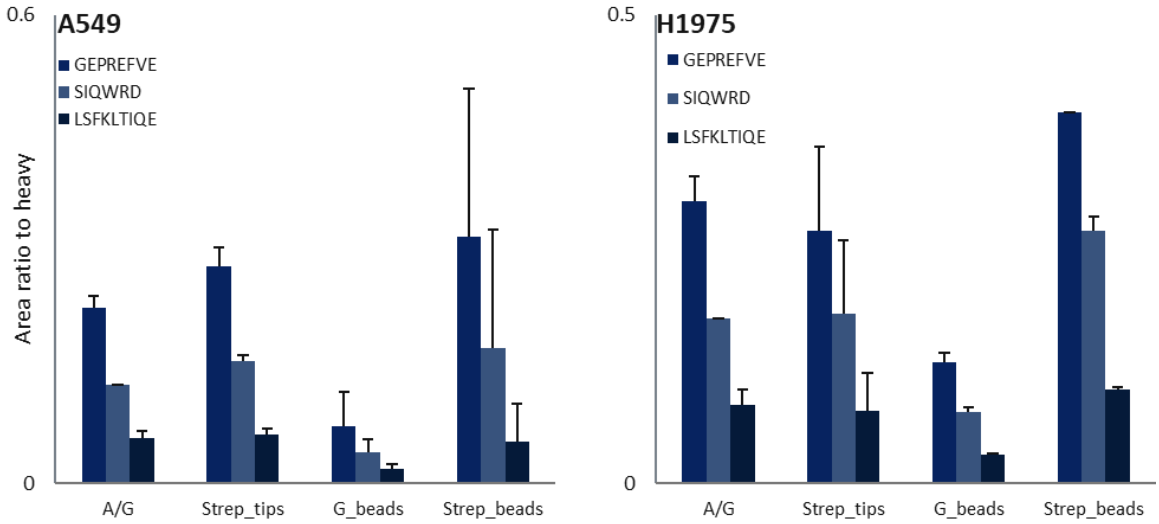


Figure I.7. Comparison of protein A/G, Streptavidin micro-column tips, protein G magnetic beads and Streptavidin magnetic beads affinity supports using anti-EGFR (clone 528) normal and biotinylated antibodies. Bars present endogenous/heavy peptide area ratio intensities of the three EGFR control peptides presented as mean ± SD variations calculated from n=2 replicates from A549 (left graph) and H1975 (right graph) cell lines. Protein A/G showed less variability between replicates.

During the development of the IP step, cross-linking of the antibody to the affinity support using dimethyl pimelimidate (DMP) crosslinker was examined to improve the overall protein enrichment and recovery. This was performed to preserve the antibody onto the support and thus to strengthen the antibody-antigen binding, which allowed usage of stringent washes afterwards for better removal of the nonimmunoaffinity-associated components. However, this assay condition did not significantly improve the protein purification and due to its time consumption was excluded from further experiments.

The final critical parameter of the protein purification step was the elution of the target from the affinity support by disrupting the antibody-antigen complex. The targeted protein can be eluted in strong acidic or basic conditions. Acidic elution buffers, containing 0.1% formic acid in water (pH=2.5), 0.4% trifluoroacetic acid in 40% acetonitrile (pH<3) or 0.1 M glycine in water (pH=2.5) were compared to buffers having 1 M ammonia or 150 mM ammonium hydroxide solution in water, both having a pH around 11. The eluates were visualized by

the SDS-PAGE method using Coomassie blue staining; the optimal band was observed with 0.4% trifluoroacetic acid in 33% acetonitrile buffer (figure I.8). Since this elution buffer was also recommended by the affinity micro-column manufacturer, it was chosen for subsequent IP performances.

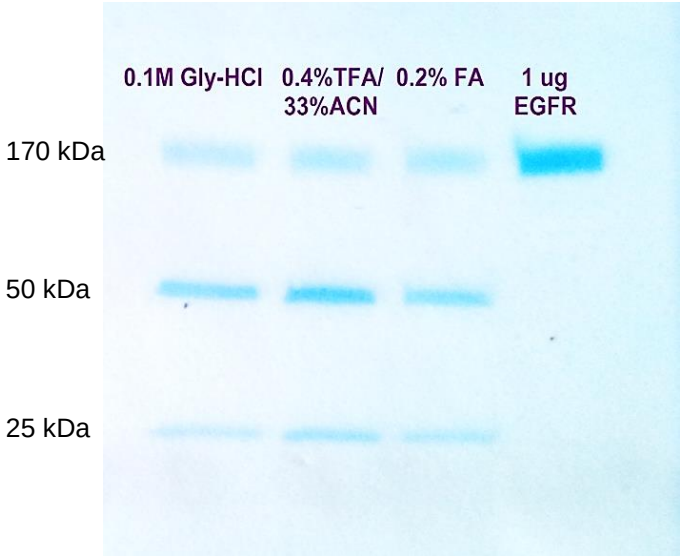


Figure I.8. Comparison of elution buffers containing 0.1 M glycine-HCl in water, 0.4% trifluoroacetic acid in 33% ACN in water and 0.2% formic acid in water for EGFR extraction; 1 µg of rEGFR was used as control. Optimal elution was obtained with the 0.4%TFA/33%ACN elution buffer.

Examination of all critical parameters of the immunopurification step resulted in the selection of a suitable antibody (supplied by Novus Biologicals) and affinity support (protein A/G micro-columns) and optimal assay settings (ammonium acetate and water serial washes and acidic elution conditions). Additionally, anti-Pan-Ras (clone RAS 10) monoclonal antibody supplied by Merck Millipore was selected for purification analysis of the RAS proteins (described in the following chapter).

Selection of signature peptides

The bottom-up mass spectrometry-based approach depends on the generation of representative peptides with sequences covering the targeted part of the protein, *i.e.*, the driver mutation, the corresponding wild-type sequence or the presence of PTMs. These peptides are obtained by digestion of the protein with suitable proteases in order to allow

their detection in the mass spectrometer. The produced peptides should have a length between 6 and 30 amino acids for simplified ionization and fragmentation during the MS analysis [101]. Furthermore, the reduction and alkylation conditions for cysteine residue removal and disulfide bond breakage prior digestion as well as the protein-to-enzyme ratio, incubation time, temperature and pH of the solvents are important for an optimal digestion step [102].

Trypsin is the most favored enzyme for protein digestion as it provides the highest sequence coverage and digestion efficiency. However, sometimes the location of the targeted alteration cannot be reached with trypsin as it would result in too short or too long peptides with decreased LC-MS performances [103]. Therefore an alternative protease had to be selected to cover the EGFR deletion and point mutations along with their wild-type counterparts. GluC endopeptidase, which cleaves after glutamic (E) and aspartic (D) acids, was selected as the alternative enzyme (figure I.9).

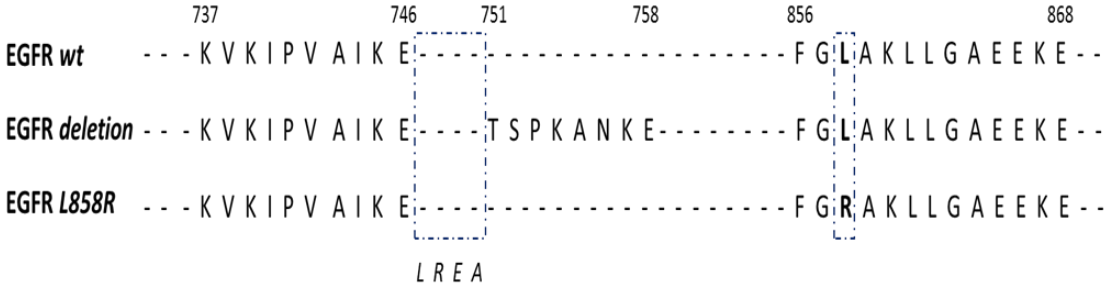


Figure I.9. EGFR signature peptide sequences covering the deletion (E746-A750) (left part of the sequence) and point (L858R) (right part of the sequence) mutation with the wild-type representatives obtained after GluC digestion. The position of both mutations is marked blue.

In addition, despite the selection of GluC endopeptidase for covering the EGFR mutation sequences, trypsin was used for generating peptides covering the EGFR tyrosine phosphorylation sites as well as the G12S KRAS mutation and the RAS family isoforms (KRas, NRas and HRas) (figure I.10), in analyses described in the following chapters.

Figure I.10. RAS signature peptide sequences covering the G12S mutation (left part of the sequence) in the wild-type representative and K, H and NRas isoforms (right part of the sequence) obtained after trypsin digestion. The position of the mutation and the isoform differences are marked in blue.

Synthetically labelled peptides

For targeted MS analysis, synthetically labelled peptides (SIL) with stable isotope labeling are used as internal standard (IS) controls to confirm the detection of the endogenous peptides by retention time and spectral matching. They can also correct possible variations of signal due to the suppression effects at the MS level and injection volumes at the LC level. Furthermore, if the exact concentration of the SIL peptides is known, the concentration of the endogenous counterpart can be calculated. These synthetized peptides with high degree of purity have the same sequences as the endogenous peptides, holding a stable isotope labeled amino acid to create an increase of the molecular mass to make them perceptible from the endogenous counterpart by mass spectrometry. When the heavy peptide is mixed with the target (endogenous peptide) in the biological sample they behave in a similar way during the LC-MS analysis and are easily distinguished during data processing due to the mass difference. Quantity of the endogenous peptide is calculated using endogenous-to-heavy peptide area ratio and the known concentration of the heavy peptide.

The earliest stage in our workflow for adding an internal standard is after the digestion step, *i.e.*, after generation of the desired peptides to be controlled. Initially, SIL peptides are spiked into the matrix without the targeted protein, to control their behavior during the LC-MS analysis and adjust the retention times to help the identification of the endogenous peptides. After spiking the IS into the sample digests, the removal of salts and other

impurities in the samples, that might interfere with the LC-MS analysis, is achieved by the solid phase extraction (SPE) method, using C18 packed columns.

The peptide sequences comprising the EGFR deletion and point mutation along with their corresponding wild-type representatives and the three control peptides for EGFR expression are presented in table I.2. Also, the peptide sequences for the G12S KRas mutation and RAS family isoforms along with the wild-type counterparts are presented in table I.3.

Table I.2. Peptide sequences covering total EGFR, targeted mutations and their wild-type counterparts.

Protein	Gene localization	Mutation	Sequence
EGFR	Exon 1 Exon 2 Exon 4	Control peptide 1 Control peptide 2 Control peptide 3	LSFLKTIQE SIQWRD GEPREFVE
	Exon 19 (deletion mutation)	*del*E746-A750	KVKIPVAIKTSPKANKE
		wild-type	KVKIPVAIKE
	Exon 21 (point mutation)	L858R (two missed cleavages)	FGRAKLLGAEEKE
		wild-type (two missed cleavages)	FGLAKLLGAEEKE
		L858R (one missed cleavage)	FGRAKLLGAEE
		wild-type (one missed cleavage)	FGLAKLLGAEE

Table I.3. Peptide sequences covering the KRas G12S mutation, its wild-type counterpart and RAS family isoforms.

Protein	Gene localization	Mutation	Sequence
KRas	Exon 2	G12S	LVVVGASGVGK
		wild-type	LVVVGAGGVGK
RAS family	Exon 4	KRas isoform	SFEDIHHYR
		NRas isoform	SFADINLYR
		HRas isoform	SFEDIHQYR

Parallel Reaction Monitoring (PRM) targeted MS analysis

Targeted MS analysis is used for precise quantification of well-defined peptide sets in complex biological samples. The targeted peptides are monitored with a high degree of sensitivity and selectivity using high resolution and accurate mass (HRAM) instruments. One such targeted method is the parallel reaction monitoring mode which simultaneously measures all the ions (precursor and fragment ion pairs) in one MS/MS scan. With this method large number of peptides can be analyzed with preferred level of sensitivity, accuracy and precision [88].

The PRM method performed on hybrid HRAM instruments resulted in precise differentiation between fragmented ions from the background signal, especially for the low-abundant peptides. The full MS/MS spectrum was acquired during the targeted analysis, available at any time. For design of the PRM method only the precursor ion *m/z,* the expected retention time, the quadrupole isolation window width, the maximum fill time and the Orbitrap resolving power are required [88]. As presented in figure 9, during the PRM analysis the predefined precursor ions are isolated and transferred into the collision cell and fragmented at a defined collision energy. Then, the resulting fragment ions are transferred to the C-trap of the instrument from where they are directed to the Orbitrap, where full MS/MS spectrum is obtained for each ion. The measured signals by this high resolution method are less subjected to interferences and therefore extensive post-acquisition data processing can be avoided.

For the targeted EGFR analysis, a 37-minute chromatographic separation time was selected as most optimal elution gradient. The main advantage of the targeted PRM method is the high resolution of the Orbitrap mass analyzer and multiplex capabilities of the C-trap and the collision cell [104]. The acquisition parameters for the full MS analysis (MS1), the resolution and maximum fill time were set at 35 000 and 200 ms, respectively, for a scanning mass range of 300 to 1500 *m/z*. The resolution for the PRM analysis (MS2) was set to 70 000, whereas the isolation window and maximum fill time were set at 1 *m/z* and 250 ms, respectively. The normalized collision energy was determined for each peptide independently. The same acquisition parameters were used for the targeted PRM analysis described in chapter II and chapter III (only difference in chromatographic separation time), as well as the phosphorylation analysis described in chapter IV (also difference in chromatographic separation time).

Method optimization

Developing a strategy for biochemical analysis of clinically relevant samples requires optimization and validation of the method prior to its implementation for routine analysis. The optimization verifies if the performance of the critical parameters is within satisfying ranges of standardized criteria. These verifications include examination of the selectivity, precision and accuracy, limits of detection and quantification, linearity range, robustness of the platform and possible interferences and matrix effects on the optimization tests [86].

Concerning the developed IP-PRM approach, the immunopurification step had the highest impact on the workflow followed by the variability from the biological material and protein extraction efficiency. Therefore, (1) the recovery after the immunopurification step was tested as an accuracy verification step, (2) the intra- and inter-day variability parameters were calculated for the method's precision, and (3) the linearity range was estimated to determine the relationship between the concentration ranges of an analyte with the corresponding signals obtained at each concentration point.

Selectivity

The ability to detect and identify an analytical target among various compounds in a complex biological matrix is considered as selectivity of a method. The capacity of the IP-PRM approach to identify the targeted proteins in a biological sample without background interference was confirmed by the successful detection of the peptides of interest covering the targeted mutations in the mass spectra. In this case, the developed methodology holds two levels of selectivity, the first coming from the monoclonal antibody used for EGFR isolation and the second from the targeted mass spectrometry analysis.

The selectivity of a method is confirmed by the detection and the identification of the targeted analyte. The capability to differentiate between signals obtained from various protein targets is one of the main advantages of the MS method as a proof for the selectivity of the instrument. The multiple signals produced, *i.e.* mass transitions for each target eluting on different retention times during the LC separations were verified by the internal standards and served as a confirmation of the MS method selectivity.

The detection and identification of the EGFR mutations after protein enrichment confirmed the selectivity of the monoclonal antibody towards the target. Moreover, the differentiation between multiple signals in the mass spectra on specific *m/z* values proved the sensitivity and selectivity of the MS and overall targeted analyses.

Recovery

A method is considered as accurate if the experimentally measured value of an analyte is close to the expected theoretical standard value. The calculated rate between these two values shows the bias of the method and regarding the IP-PRM approach it can be assessed by estimation of the immunopurification step recovery. For the recovery calculation usually three to five experimental replicates are needed to be compared to known reference standards.

Recombinant EGFR (rEGFR) protein spiked into H3255 lung adenocarcinoma cell line extracts was used for calculation of the IP recovery. Namely, five individual rEGFR replicates in 100 mM phosphate buffer were digested to serve as reference (standard set, n=5). Five replicates were prepared by spiking rEGFR into five individual H3255 cell lysates representing the biological matrix (set 1, n=5). Five different individual replicates of the H3255 adenocarcinoma cell line were used in the recovery calculations to deduct the endogenous EGFR amount (set 2, n=5). Sample sets 1 and 2 were subjected to the previously described IP-PRM protocol, whereas the standard set was only analyzed by targeted PRM.

For calculation of the recovery, the obtained signals of the three EGFR control peptides were used in regard to the known amount of injected rEGFR (200 fmol of protein per injection) and results are presented in table I.4. For each peptide, the recovery was calculated from the mean value of n=5 replicates. The recovery percentages for each control peptide were calculated according to the equations presented below:

Recovery equation

$$\frac{((\text{set 1}) - (\text{set 2}))}{(\text{standard set})} \times 100\% \quad \ldots\ldots\ldots \quad (1)$$

Table I.4. Calculated recoveries of rEGFR spiked in H3255 cell lysate and the matrix effect

Control peptides (n=5)	GEPREFVE	SIQWRD	LSFLKTIQE	Average
rEGFR spiked in H3255 cells	32±0.02%	14±0.01%	12±0.01%	20±0.02%

Results demonstrated that only about 20% (mean value of the three control peptides) of EGFR was recovered during the purification step. The average value was calculated as identical results were expected for all the three control peptides since they represent the total EGFR expression in the cells.

To check the calculated recovery, EGFR extracts from A431 cell lysates were enriched and all the IP steps – including the flow-through after protein capture, ammonium acetate, PBS and water washes, and the elution step – were analyzed by WB, using antibody against the total EGFR and total cell lysate as control. The gel bands in figure I.11 showed the lower recovery, with minimal losses of EGFR during the washing steps. Most of the EGFR remained uncaptured (flow-through) probably due to the lower amount of antibody used for protein enrichment or the impact of the detergent on the binding efficiency.

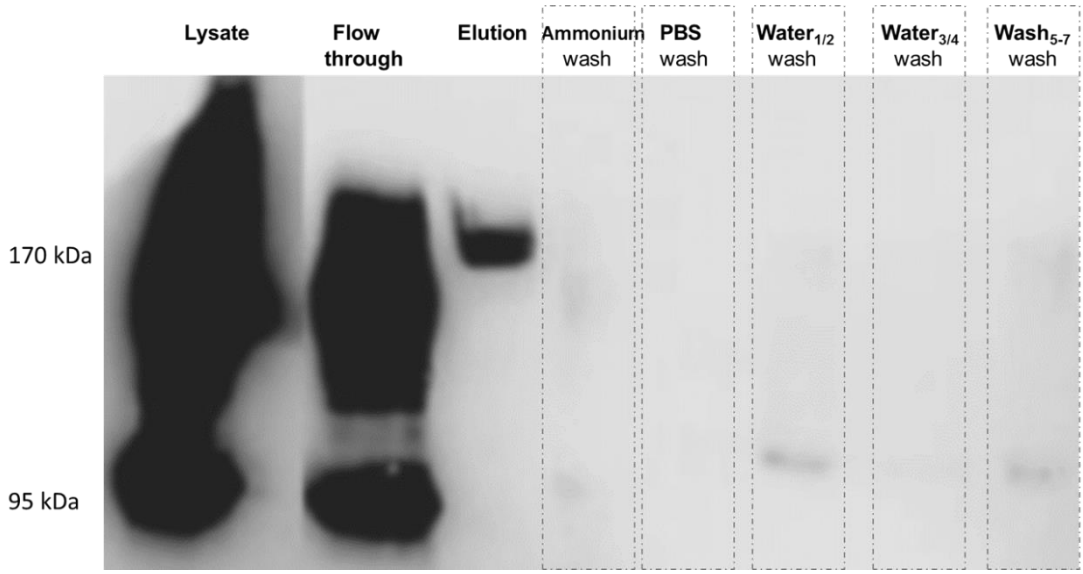


Figure I.11. WB analysis of the IP steps. A431 cell lysates were immunopurified using anti-EGFR monoclonal antibody to check the EGFR recovery. The total lysate, flow-through after capture, elution, ammonium acetate, PBS and combined water washes are presented in each well of the gel. The enrichment and recovery of EGFR was low (3rd well) with minor losses during ammonium acetate wash and first and third set of water washes. 170 kDa is a molecular mass of total EGFR, whereas 95 kDa is for the external part of the protein.

Precision

The precision of a method characterizes the variability between repeated measurements of the same sample quantity and is usually presented as a coefficient of variation (CV). The measurement variations may be observed within a day (intra-day precision or repeatability) and between measurements in different days (inter-day repeatability). When replicated values are close together the method is considered as highly precise, showing CV values below 20%.

To calculate the repeatability of the IP-PRM workflow, H3255 cell extracts, were analyzed in triplicate under the above described conditions. The replicates were initially prepared from the same pool of cells which were aliquoted and stored at -80°C, to avoid any variability coming from the biological material. For intra- and inter-day precision measurements the three EGFR control peptides were monitored by PRM and the obtained endogenous/synthetic peptide area ratios were used to calculate the CVs. As mentioned before, similar results are expected for each control peptide representing the EGFR expression in the cells and thus the final results are presented as an average CV of the three peptides (tables I.5 and I.6). The percentages in table I.5 were calculated as the mean of n=3 replicates prepared and measured on the same day, for each peptide, whereas CVs presented in table I.6 were calculated as the mean of n=3 replicates prepared and measured on different days, to monitor the variability between the instrument performance and overall repeatability of the workflow.

Table I.5. Calculated CV values for intra-day precision of the IP-PRM method (n=3).

Peptide	GEPREFVE	SIQWRD	LSFLKTIQE
Average CV	15±0.04%	14±0.05%	9±0.02%

Table I.6. Calculated CV values for inter-day precision of the IP-PRM method (n=3).

Peptide/Day	GEPREFVE	SIQWRD	LSFLKTIQE
Average CV	13±0.01%	14±0.005%	11±0.05%

The developed workflow, consisting of six independent steps, demonstrated variability of less than 15% in both measurements (intra- and inter-day).

Linearity range

In biochemical assays, linearity is defined as the ability of a method to generate results proportional to an analyte concentration in a given range, with corresponding precision and accuracy. The linearity should be established for a defined working range, using at least five different concentration values of the targeted analyte. For an LC-MS analysis, the linearity range depends on the analyzed compound, where endogenous/synthetic peptide area ratios are obtained for each analyte concentration.

To estimate the linearity range of the IP-PRM method, rEGFR in ten diverse concentrations was spiked in ten individual cell extracts from the A549 cell line (biological matrix with low levels of EGFRwt expression). The rEGFR concentration range was between 0 and 100 ng/µL corresponding to 0 and 1 pmol protein injected, respectively. Once more, the three EGFR control peptides were observed, and the signal intensities obtained for each peptide are presented in figure I.12.

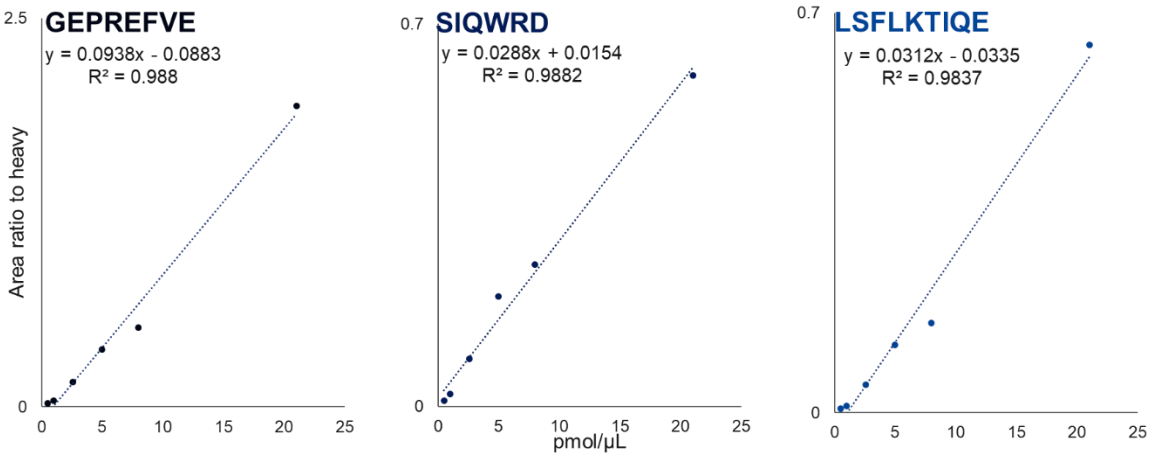


Figure I.12. Linearity of IP-PRM. Estimated working range presented as intensities signals for each control peptide obtained for 0 to 100 ng/µL analyzed rEGFR.

The results demonstrated acceptable linearity of the workflow within the given range, with a coefficient of correlation > 0.98.

Method comparison

After development and optimization of the IP-PRM approach, the next step was to evaluate its performance. Comparison of the method – covering generation of results, acceptable accuracy and precision – to commercially available or routinely used techniques in other laboratories can serve as a validation step. Confirmation of the performance depends on the variation between the obtained results from both approaches.

A membrane fractionation technique was chosen for comparison to the immunopurification step of our workflow, as it is a widely used method for membrane protein isolation from cell matrices. Membrane fractionation as a simple and fast approach requires modest equipment and reagents; membrane proteins are isolated from the other subcellular compartments by ultracentrifugation at 100 000 x g. Therefore, in step 3 of our workflow in figure I.1, immunopurification was substituted by membrane fractionation, to isolate EGFR from the four lung adenocarcinoma cell lines. Membrane fractions were pelleted by ultracentrifugation and subsequently digested with GluC endopeptidase and analyzed by PRM. The acquired signal intensities (endogenous/synthetic peptide area ratios) for the three control peptides covering the EGFR expression within each cell line were compared to the intensities obtained after the immunopurification analysis of EGFR. The graphs in figure I.13 showed similar performances of both methods, immunopurification of EGFR (right graph) and membrane fractionation (left graph), after comparison of the obtained signal intensities from the four cell lines. Further, the repeatability of the membrane fractionation step was assessed by comparing the intra- and inter-day variability. Following the settings described in the *Precision* section from this chapter, calculated CVs showed greater variability of the membrane fractionation method within a day and between different investigations (CVs between 12 and 34%) compared to the IP-PRM CVs. The results presented in tables I.8 and I.9 evaluated the performance of the IP-PRM method.

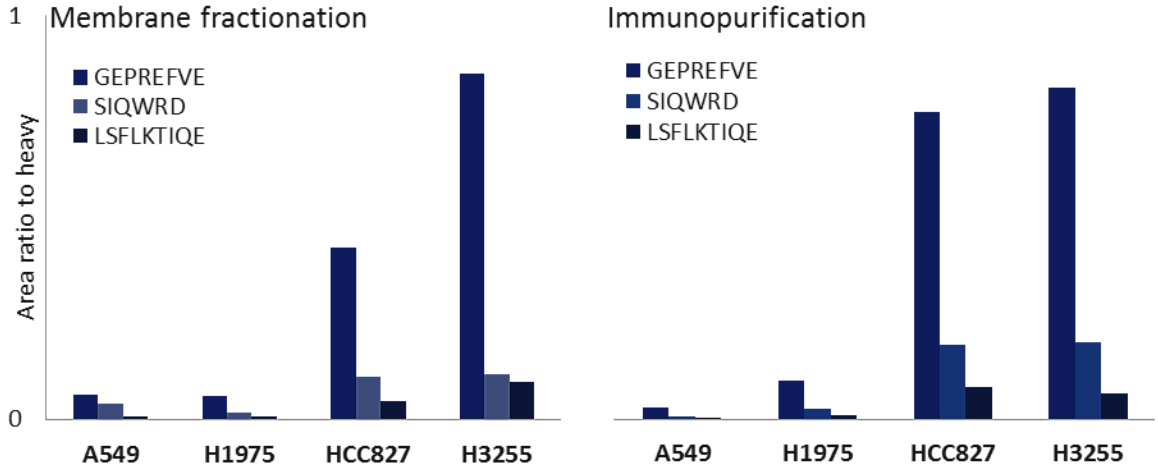


Figure I.13. EGFR signal intensities obtained after membrane fractionation (left) or immunopurification (right). Bars present endogenous/heavy peptide area ratio intensities of the three EGFR control peptides presented as the mean calculated from n=2 complete replicates.

Table I.8. Calculated CV values for intra-day precision of the MF-PRM method (n=3).

Peptide	GEPREFVE	SIQWRD	LSFLKTIQE
Average CV	19±0.07%	12±0.08%	31±0.2%

Table I.9. Calculated CV values for inter-day precision of the MF-PRM method (n=3).

Peptide	GEPREFVE	SIQWRD	LSFLKTIQE
Average CV	24±0.13%	22±0.08%	34±0.21%

DISCUSSION

Protein isoforms as products of altered genes, characteristic for different cancer types, exist in various abundance and forms in cancer cells, and as drug targets their unambiguous interrogation is important for understanding disease processes and pathological events [105, 106]. Therefore, effective and multiplexed assays for screening patients carrying oncogenic “driver” mutations are essential for their detection and quantification at a nano-gram scale. Considering these requirements, a stepwise workflow was developed and optimized for “driver” mutation investigation by combining immunoaffinity purification with targeted PRM analysis.

Immunohistochemistry (IHC) [107, 108] and various protein arrays [109, 110] using mutation-specific antibodies are widely used for protein analysis with clinical relevance. However, these specific antibodies can be limited regarding the target, making these approaches not suitable for analysis of specific “driver” mutations. On the other hand, the enzyme-linked immunosorbent assay (ELISA) that uses antibodies for protein recognition and capture is applied in routine clinical protein analyses, but cannot differentiate between mutations, PTMs or other variants [111]. On the contrary, the described single-step immunoaffinity purification using pan-antibodies enriched the endogenous protein preserving its modifications and allowed subsequent targeted analysis with decreased complexity, minimized ionization suppression and shorter gradient chromatographic separation [95]. Furthermore, isolation of the total protein permitted examination of mutations, isoforms and PTMs in a single PRM analysis, without additional enrichment, depletion or fractionation step [82, 112]. The ability to clearly distinguish between signals at different *m/z* values – representing unique protein signatures – in a qualitative and quantitative manner, presents a valuable resource for studying disease-specific protein variations with reference to assist clinicians in personalized therapy selections.

There were some limitations in our method development and optimization. The primary challenge was availability of high-quality antibody. Certain antibody parameters – as antibody type, purity, host, immunogen, isotype, method of application and validation – can increase the level of non-specific binding, also if the antibody is not compatible with the affinity support, low recovery of the targeted protein will be achieved [113]. Therefore, tested antibodies were chosen according to IgG isotype (IgG_1 or IgG_2 isotype as most compatible with protein A/G) and immunogen (purified EGFR from human cells). Second, the

composition and compatibility of the lysis buffer, designed for protein solubilization and extraction, decreased the accessibility of the protein binding to the antibody by cross-reacting with the antibody or the affinity support and increasing the non-specific binding. An additional issue is the impact of the detergent on the mass spectrometry analysis [114]. All detergents were selected due to their compatibility with the MS instrument and they were eluted from the LC-column at the end, thus not affecting the signals of the targets. Third, IP assay conditions (pH, temperature, reagent composition, storage etc.) affecting the formation of antibody-antigen complex were tested to avoid generation of poor quality spectra (e.g. pH=8.2 is optimal assay condition for protein A affinity support compared to protein A/G widen range (pH=4-9) information obtained from the manufacturer). Finally, the quality and purity of the internal standard can restrict the identification and quantification of the targeted protein if their signals do not overlap in the spectra at given retention time [88, 115]. Nonetheless, the majority of these limitations were outperformed – selection of suitable antibody, affinity support, protease and internal standard controls – enabling EGFR detection and identification from complex biological samples.

The combination of immunoaffinity capture and targeted PRM analysis offers several benefits regarding the implementation of the approach in routine analysis. First, the developed method holds two levels of selectivity for unambiguous detection of oncoproteins [116]. EGFR was identified in each cell line, where due to the reduced background complexity chromatographic separations of around 40 minutes were applied. Second, the IP-PRM approach demonstrated repeatability of 15% variation between analyses [117], accepted CV since the workflow was developed from six independent steps, each individually controlled and optimized. Comparison with the membrane fractionation method – isolation of membrane proteins with low purity and possible contaminations of other cellular compartments - confirmed the precision of the IP-PRM method. Third, the method showed acceptable correlation between the different amounts of injected EGFR and recovered MS signal for each amount. Linearity range between 0 and 100 ng/µL demonstrated that EGFR can be detected in concentrations less than 1 ng/µL verifying the high-throughput capability of the MS analyses [118]. On the contrary, the main drawback of this study was the recovery rate after EGFR immunopurification. The 20% recovery was confirmed by WB analysis of purified rEGFR in PBS buffer and endogenous EGFR from A431 cells. The matrix effect had the greatest impact on the amount of recovered EGFR,

even though the lysis buffer composition was suitable for the MS instrument and extracted EGFR the most.

High resolution accurate mass spectrometry analysis has been introduced for targeted quantitative analysis of therapeutic relevant targets in combination with immunoaffinity purification [82, 118, 119]. Moreover, the PRM method enhances the detection of targeted "driver" mutations in a high-throughput and sensitive manner; merged with the IP method, due to the decreased sample complexity, also yielded a faster analysis. The estimated total turnaround time including the complete sample preparation with overnight digestion, was about 5 days, if a multiplexed 96 well-plate format is used, *i.e.,* 96 different patients can be analyzed simultaneously [118]. Additionally, this approach is suitable for analysis of diverse samples, like plasma, urine, small biopsy aspirates, tissue, etc. [95, 120, 121] and thus being valuable for patient stratification carrying specific mutations eligible for suitable targeted therapies.

MATERIALS AND METHODS

Chemicals and Reagents

A549, H1975, HCC827, H3255 and A421 cancer cell lines were a gift from the Laboratory of Experimental Haemato-Oncology (LHCE) of the Luxembourg Institute of Health (LIH). MSIA® Disposable Automation Research Tips with protein A, protein G and protein A/G (Cat.No. 991PRT15) and with Streptavidin (Cat. No. 991STR11) along with the EGFR antibody biotin conjugated (528) (Cat. No. MA5-12872) were obtained from Thermo Fisher Scientific BVBA. Streptavidin magnetic beads (Cat. No. 88816) were purchased from Pierce. Protein G magnetic beads (Cat. No. 28-9440-08) were obtained from GE Healthcare Life Sciences. Monoclonal EGFR/ErbB1 antibody (clone 528) (Cat. No. NB110-5846) was purchased from Novus Biologicals. Anti-EGFR clone 528 (Azide free) (Cat. No. MABF119) was purchased from Millipore (MERCK). Active human EGFR recombinant protein (95 kDa external part) (Cat. No. ab155726) was obtained from Abcam. Endo-Glu-C (Staphylococcus protease V8) was obtained from Worthington. Stable isotopically labelled peptides were synthesized by Thermo Fisher Scientific. The western blot system, with all reagents and

materials were obtained from Invitrogen. All other reagents were obtained from Sigma Aldrich.

Sample preparation step

Biological material: A549 lung adenocarcinoma cells were grown in Dulbecco's modified Eagle's medium (DMEM/F12) (Lonza) supplemented with 10% (v/v) fetal bovine serum (FBS) (Life Technologies) and 1% (v/v) penicillin/streptomycin mixture (Lonza). H1975, HCC827 and H3255 lung cancer cells were grown in RPMI-1640 medium (Lonza) supplemented with 10% (v/v) FBS and 1% (v/v) Pen/Strep mix. A431 epithelial cancer cells were grown in DMEM high glucose, pyruvate medium (Life Technologies) supplemented with 10% (v/v) FBS and 1% (v/v) Pen/Strep mix. Cells were incubated at 37°C in 95% humidity atmosphere and 5% CO_2 and grown to complete confluence in T-75 or T-175 flasks. The confluent cells were collected from the flasks by washing with PBS buffer (Life Technologies), followed by incubation with 0.02% (w/v) ethylenediaminetetraacetic acid (Lonza) at 37°C for 15 minutes for cell detachment. The number of cells was estimated using Countess™ automated cell counter (Invitrogen) or manually using Bruker hemocytometer (Sigma). After centrifugation at 300 xg for 10 minutes, the supernatant was aspirated and the cell pellets were used immediately or stored at -80°C.

Cell lysis and protein extraction: Cell pellets were re-suspended in 1 mL lysis buffer (50 mM Tris-HCl pH=7-8, 150 mM NaCl, 1 mM Na_3VO_4 and 1% detergent (DDM)) containing 90 µL of protease and phosphatase inhibitors cocktail (Roche). Five to ten short sonication cycles for lysis and protein extraction were performed on ice or three freeze/thaw cycles (-80°/25°C) were used, followed by centrifugation at 20 000 xg at 4°C for 30 minutes and collection of the supernatant.

Immunopurification: Protein extracts were subjected to protein enrichment using 2 µg of mAb previously loaded on the micro-columns. The immunopurification was performed by repeated aspiration/dispense cycles of the cell extracts through the micro-column for protein binding to the antibody onto automated liquid handler Versette® working station (Thermo). After capturing the protein, the micro-columns were washed once with 2M ammonium acetate buffer (pH= 8) and 10 mM PBS (pH=7.6) buffer and additionally 7 times with short

water wash aspiration/dispense cycles. EGFR was eluted form the micro-column in 0.4% TFA/40% ACN elution buffer (pH=2.5) and eluates were vacuum dried.

Digestion, IS spiking and desalting: Dried samples were re-suspended in 30 µL of sodium phosphate buffer (pH=7.8) and reduced with 50 mM DTT for 45 minutes at 50°C and alkylated with 150 mM IAM for 30 minutes in the dark. 0.04 µg of GluC protease was added to each sample for overnight digestion at 37°C. The next day, samples were spiked with heavy labelled peptides, desalted onto solid phase extraction Sep-PakC18 cartridges (Waters), vacuum dried and re-suspended in 25 µL of 0.1% formic acid in water for LC-PRM analysis.

Western blot analysis: Concentrations of the protein extractes were determined by Qubit™ 2.0 Fluorometer with the Qubit™ Protein Assay Kit (Thermofisher Scientific Inc.). 20 µg of total protein extracts were used for sodium dodecyl sulfate polyacrylamide gel electrophoresis separation followed by their membrane tranfer using iBlot Dry Blotting System according to the manufacturer instructions (Invitrogen, Life Technologies). The blots were then incubated with EGF Receptor (D38B1) XP Rabbit mAb (#4267, Cell Signalling Technology) and Peroxidase AffiniPure Goat Anti-Rabbit IgG (H+L) pAb for 4 hours at room temperature. EGFR bands were detected by chemoluminiscence substrate SuperSignal™ West Pico PLUS (Thermo Pierce) and ImageQuant LAS 4000 system (GE Healthcare, United Kingdom).

Membrane fractionation: After cell detachment from the flasks, cells were pelleted at 500 x g to remove the media. Then, cell pellets were re-suspended in sucrose buffer and subjected to three cycles of ultracentrifugation (Beckman Coulter) at 100 000 x g, to isolate the membrane sub-cellular fraction as a pellet and remove all the other sub-cellular parts present in the supernatant. After obtaining a pellet that contains EGFR, the pellet was re-suspended in digestion buffer and undergo proteolysis.

LC-MS targeted analysis

LC separation: The chromatographic separations were done on a Dionex Ultimate 3000 RSLC chromatography system, operating in a column switching setup. The mobile phase A consisted of 0.1% formic acid in water and the mobile phase B of 0.1% formic acid in acetonitrile. The loading phase of the samples was composed of 0.05% trifluoroacetic acid

and 1% acetonitrile in water. Samples were injected and loaded onto a trap column (75 µm x 2 cm, C18 pepmap 100, 3 µm) at 1 µL/min or 5 µL/min, followed by elution onto an analytical column (75 µm x 15 cm, C18 pepmap 100, 2 µm) with 300 nL/min flow rate. Separation was done by a linear gradient starting from 2% to 90% B in 37 min.

PRM analysis: The PRM analyses were performed on a QExactive Plus (Thermo Scientific) mass spectrometer equipped with an EASY-spray ion source. The PRM method was performed with a quadrupole isolation window of 1 m/z units, an automatic gain control target of 1e6 ions, a maximum fill time of 250ms and an orbitrap resolving power of 70000 at 200 *m/z*. Collision energy was optimized for each precursor. The duration of the scheduled time windows for each pair of endogenous and heavy labelled peptides was set to 2 min.

Data processing: Fragment ion chromatograms were extracted from the MS raw data and processed using the Skyline package software version 3.7.0.11317. Fragment ions were selected according to the accuracy of the mass measurement and the co-elution and corresponding fragment patterns between the endogenous and isotopically labeled standards. For each peptide, the ratios between the sum of the fragments of the endogenous peptides and the labelled ones were calculated.

Chapter II

Targeted PRM analysis

BACKGROUND

Genomic alterations such as gene mutations, copy number variations and/or mutant allele specific imbalance (MASI) are well described in literature, whereas the consequences of these changes occurring at protein level are poorly understood [4]. Estimation of the relative expression levels of “driver” mutation versus the wild-type in oncoproteins used as drug targets is of a great importance, especially considering the MASI genomic changes where the mutant allele is amplified and/or the wild-type allele is deleted [122]. Various “driver” mutations have impact on targeted therapies, but it would be inefficient and time-consuming to use individual and specific drug tests for each candidate. Instead, multiple targets known to harbor “driver” mutations with predictive value can be observed using a targeted mass spectrometry approach [81].

The study described in this chapter presents a modified version of the previously published IP-PRM approach [95]. The formerly developed methodology was used for targeted PRM analysis of KRas and EGFR mutations in lung adenocarcinoma cell lines and tissue and of serum amyloid A (SAA) isoforms in lung cancer plasma samples. The basis of this work was the application of a fast-LC chromatographic separation prior to the targeted PRM analysis, where a short (7 min) gradient time could be applied due to the decreased sample complexity obtained by protein immunopurification. This IP-fastLC-PRM approach was able to identify all the targets in less than 24h with an additional overall sample preparation time of 24-36h. It was concluded that with further usage of a 96-well plate format and a shorter digestion procedure this method can provide fast results for clinicians who need to make decisions regarding therapeutic strategies based on the patients’ mutation status.

Here, the analytical platform combining the protein immunopurification with the targeted PRM analysis (described and optimized in Chapter I) was used for identification and quantification of RAS family isoforms and EGFR mutations in the five cancer cell lines already described in Chapter I. Compared to the aforementioned approach, where a multichannel pipettor was used for protein immunopurification, here the IP step was automatized using the Versette® liquid handling station working in a 96-well plate format. This high-throughput and selective detection platform unambiguously distinguished between mutated and wild-type sequences from ng/µL samples using a 37-minute-long chromatographic separation. The chromatographic separation time was increased

from 7 to 37 minutes due to the use of a different LC instrument. However, it was still much shorter than the standard 60 to 90 minute gradient time.

Furthermore, this set-up allowed us to quantify at protein level a set of “driver” mutations and later post-translational modifications with high sensitivity and selectivity. The results obtained by this IP-PRM approach could support and facilitate the verification of the tumor’s driver mutation heterogeneity, which is usually mainly based on genomic analyses.

RESULTS

Identification and quantification of Ras family isoforms

The Ras family consists of three isoforms with GTPase activity: The Kirsten rat sarcoma virus (KRas), the neuroblastoma RAS viral oncogene homolog (NRas) and the Harvey sarcoma virus (HRas). The sequence in the nucleotide binding region (1-119 amino acids) of these isoforms is identical, indicating equal binding characteristics for all three isoforms. The G12S mutation (one of the most common KRas substitution mutations) can be distinguished by the presence of ^{6}LVVVGASGVGK^{16} (mutated representative peptide) versus ^{6}LVVVGAGGVGK^{16} (wild-type counterpart sequence), while the three isoforms can be discriminated by the signature peptides ^{89}SFEDIHHYR^{97} (KRas), ^{89}SFADINLYR^{97} (NRas) and ^{89}SFEDIHQYR^{97} (HRas) (sequences presented in figure 3 and I.10).

Ras proteins were purified from the lung adenocarcinoma cell lines using a monoclonal anti-Pan-Ras antibody and subjected to trypsin digestion for signature peptide generation. Results showed that the G12S mutation was only harbored in the A549 cells, while the other cells expressed the wild-type KRas. As presented in figure II.1A, the G12S mutation was only detected in the A549 cells, at 62% mutation rate, describing this cell line as heterozygote. All the other cells were characterized as homozygotes, expressing only the wild-type KRas. Furthermore, all three Ras isoforms were identified in all the adenocarcinoma cells, although at diverse frequencies (figure II.1B). The KRas, NRas and HRas isoforms were expressed in the A549 cells at a 51%, 40% and 8% frequency, respectively. In the H1975 cell line the same three isoforms were observed at 44%, 46% and 10% expression rate. The frequency of the K, N and H isoforms was 42%, 54% and 4% in the HCC827 cells, respectively, while in the H3255 cell line the isoforms were expressed at a 47%, 34% and 19%, correspondingly. The observed expression pattern of

the three isoforms in the A549 and H3255 cells was KRas>NRas>HRas, whereas in the H1975 and HCC827 cell lines the isoform expression was NRas>KRas>HRas. HRas was always the least expressed isoform in all the cell lines. Additionally, the representative LC-PRM chromatogram profiles of the signature peptides for the KRas and Ras family isoforms in each of the cell lines are presented in figure II.2.

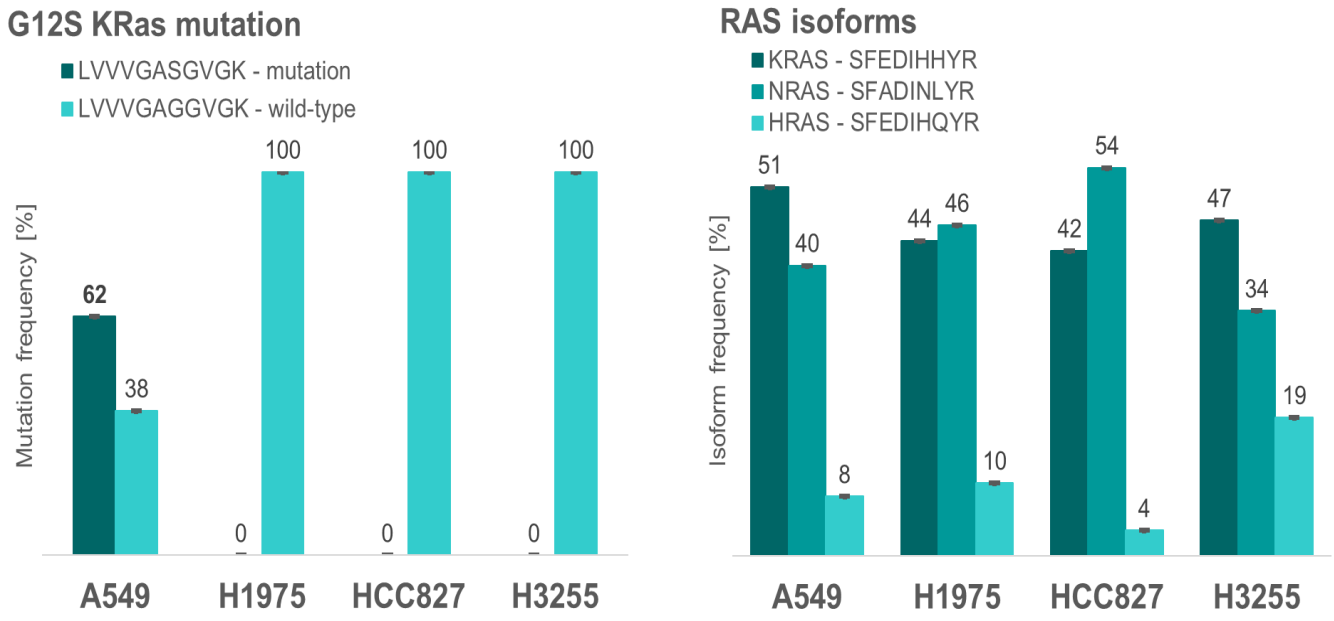


Figure II.1. KRas G12S mutation rate (left) and KRas, NRas and HRas isoform frequency (right) in four different lung adenocarcinoma cell lines. The G12S mutation was only detected in the A549 cells, while the isoform frequency was estimated among all the cells. Bars represent endogenous/heavy modified and unmodified peptide area ratio intensities presented as the mean±SD calculated from n=2 replicates.

Identification and quantification of EGFR deletion and point mutations

The EGFR deletion (EGFR$^{del E746\text{-}A750}$) and point mutation (EGFR^{L858R}) are the most frequent mutations occurring in the tyrosine kinase domain of the protein (over 90% frequency). The deletion mutation is harbored only in the HCC827 cell line and can be detected by the presence of the ^{737}KVKIPVAIKTSPKANKE^{758} signature peptide. The H1975 and H3255 cells harbor the leucine-to-arginine substitution mutation, represented by the ^{856}FGRAKLLGAEEKE^{868} signature peptide. The A549 and A431 cell lines are only

expressing the wild-type EGFR. The representative peptide for the wild-type counterpart of the deletion mutation is the ^{737}KVKIPVAIKE^{746}, whereas the wild-type of the L858R mutations is represented by the ^{856}FGLAKLLGAEEKE^{868} peptide sequence.

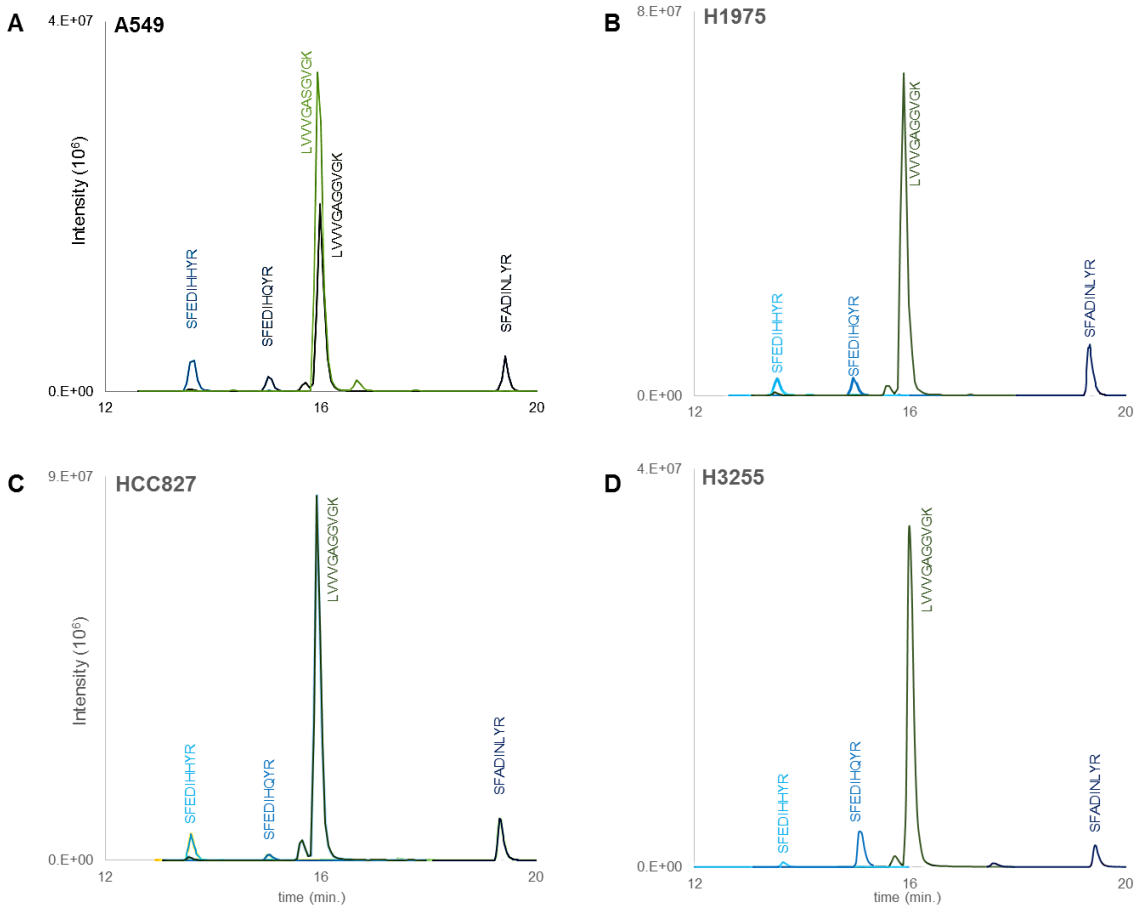


Figure II.2. LC-MS chromatograms of the signature peptides of the Ras family isoforms, including G12S mutation and wild-type KRas, K, N and HRas, from (A) A549, (B) H1975,(C) HCC827 and (D) H3255cell lines.
The G12S mutation is represented via LVVVGASGVGK peptide, present only in A549 cells; the corresponding wild-type sequence is LVVVGAGGVGK.
Representative peptides for the three isoforms are: SFEDIHHYR for KRas, SFEDIHQYR for HRas and SFADINLYR for NRas isoform.

The described IP-PRM approach was applied for identification and quantification of these mutations in the previously defined cancer cells. In this analysis, the A431 epidermoid

cancer cell line was included as a control due to the overexpression of EGFRwt. As presented in figure II.3A, the L858R point mutation was identified only in the H1975 and H3255 cells via FGRAKLLGAEEKE, the representative peptide. The quantitative PRM analysis described these two cells as heterozygotes, estimating a mutation rate of 55% in the H1975 cells and of 91% in the H3255 cell line. The wild-type representative peptide, FGLAKLLGAEEKE, was identified in all the five cancer cell lines. On the other hand, the deletion mutation, represented by the KVKIPVAIKTSPKANKE peptide, was detected at 92% mutation frequency only in the HCC827 cell line, whereas KVKIPVAIKE, the wild-type counterpart was identified in all the cells (figure II.3B). The LC-PRM profiles of lung adenocarcinoma cell line regarding the EGFR mutation status are presented in figure II.4.

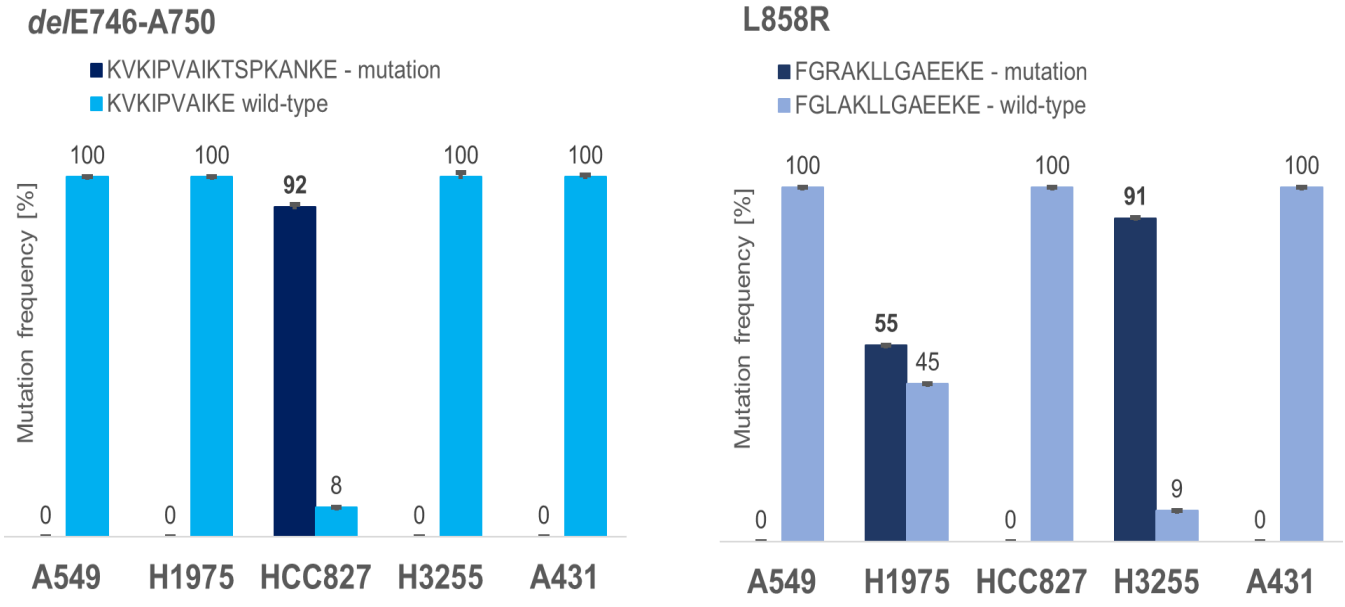


Figure II.3. Estimation of the EGFR *del*E746-A750 (left) and L858R (right) mutations in the five cancer cell lines. The deletion mutation was detected and quantified only in the HCC827 cell line, while the point mutation was identified and quantified in the H1975 and H3255 cells. Bars represent endogenous/heavy peptide area ratio intensities of the modified and unmodified peptides presented as the mean±SD calculated from n=3 replicates.

The application of the developed and optimized targeted IP-PRM approach on the five cancer cell lines expressing different Ras family isoforms and EGFR mutations resulted in the clear identification and quantification of all the targets. This was possible due to the

decreased sample complexity and the increased overall sensitivity and selectivity, obtained using 37-minute-long chromatographic separations.

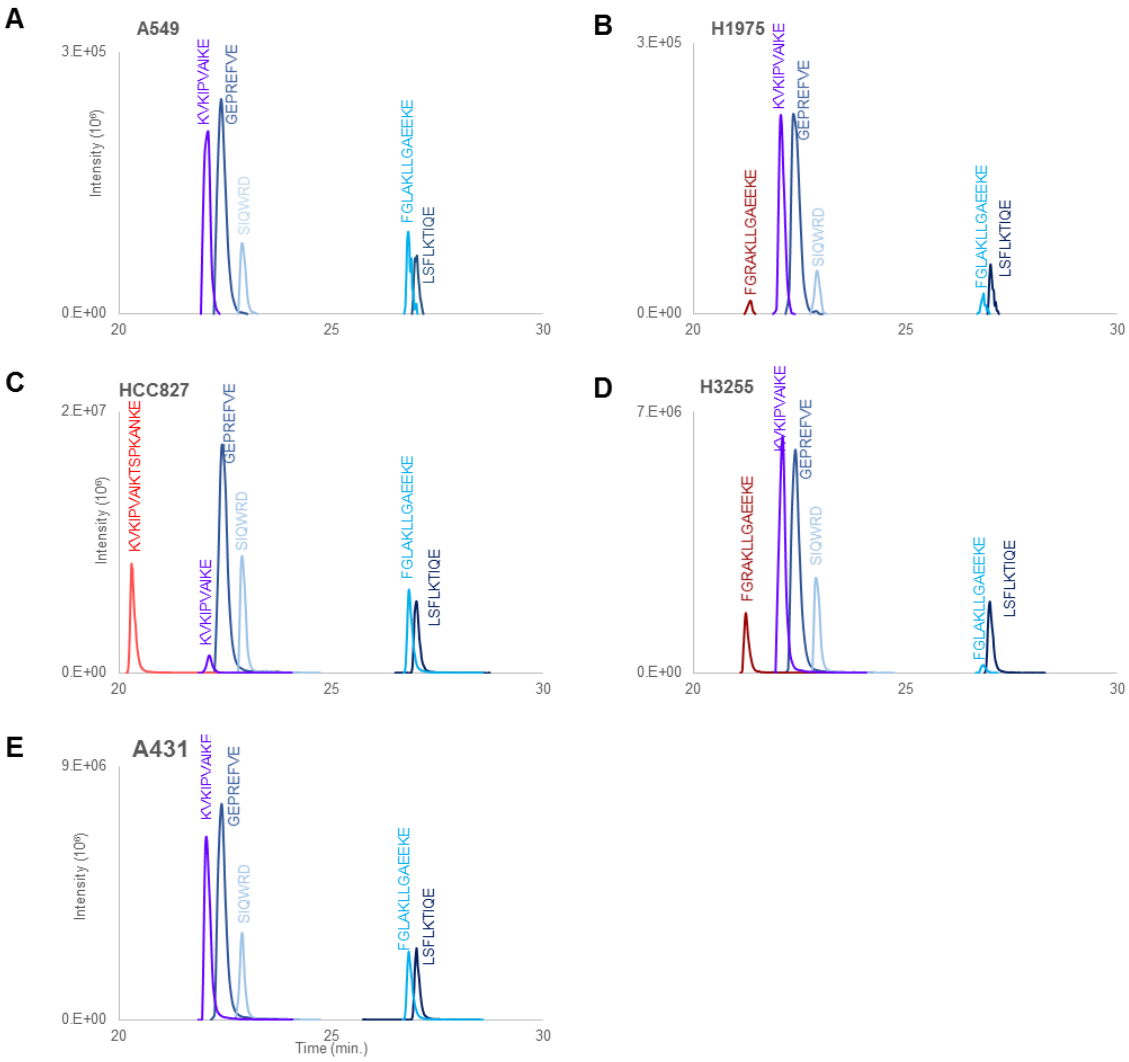


Figure II.4. LC-MS chromatograms of the signature peptides of EGFR mutations, including delE746-A750 mutation, L858R mutation and the corresponding wild-type peptides, from (A) A549, (B) H1975, (C) HCC827, (D) H3255 and (E) A431 cell lines.
The delE746-A750 mutation is represented via KVKIPVAIKTSPKANKE peptide, present only in HCC827 cells; the corresponding wild-type sequence is KVKIPVAIKE.
The L858R mutation is represented via FGLAKLLGAEEKE present only in H1975 and H3255 cell lines; the corresponding wild-type sequence is FGRAKLLGAEEKEE.
The GEPREFVE, SIQWRD and LSFLKTIQE represent control peptides for total EGFR expression.

DISCUSSION

The aim of this work was to develop and implement an analytical platform for unambiguous differentiation of oncoprotein isoforms known to harbor “driver” mutations. EGFR and KRas mutations were selected to serve as proof-of-principle due to their clinical predictive value for various cancer types and the various existing targeted therapies that could be used for testing to obtain new insights in the cancer disease.

The commonly used mass spectrometry-based approach is a shotgun proteomics method where thousands of proteins and/or peptides can be identified. This approach is mostly applied in discovery analyses for identification of new potential biomarkers with diagnostic, prognostic, predictive or therapeutic value [123]. Some of the difficulties of this approach are the long chromatographic gradient separations (usually 90 to 180 minutes, depending on the sample complexity) and the lower sensitivity and selectivity of the analysis due to the competition of multiple abundant ions that suppress the detection of the low abundant ones [124]. Furthermore, the quantitative measurements using the shotgun approach of identified proteins (and peptides) under various biological conditions have decreased accuracy as a result of the presence of multiple targets [125]. On the contrary, with the targeted MS-based analysis, especially working in parallel reaction monitoring mode, a selection of subset of targets can be observed in a qualitative and quantitative manner on high resolution and accurate mass instruments [88]. With the targeted approach short gradient times can also be applied leading to the analysis of numerous targets in short time.

The mutation status of the KRas oncoprotein as well as the other two Ras family isoforms, NRas and HRas, can provide diagnostic, prognostic and predictive insights for various cancer types, such as colorectal, pancreatic and non-small cell lung cancer. KRas is involved in the three main downstream signaling pathways of EGFR (MAPK/ERK, JAK/STAT and PI3K) as a signal transducer, thus the estimation of its mutation status is quite important for the selection of targeted therapies [126]. For targeted MS-based analysis, trypsin was chosen to generate signature peptides covering the G12S KRas mutation and representative peptides to distinguish between the three Ras family isoforms. The signature peptide for the G12S mutation is located between the 6 and 16 amino acid in the protein sequence, whereas the discrimination between the isoforms occurs in positions 89 to 97. The peptide identification was confirmed by the synthetically labeled peptides used as internal standard controls, which co-eluted with the endogenous peptides and showed

similar fragmentation patterns in the MS spectra. The mutation was only detected in the A549 adenocarcinoma cells at 62% mutation frequency together with the wild-type counterpart (38% expression rate). The identification of the KRas G12S mutation in the cell line that harbors wild-type EGFR supported the known mutual exclusivity of these two mutations [127, 128]. In lung cancer, the KRas and EGFR mutations rarely or never occur in the same tumor probably due to the equivalent contribution of these two proteins (as gene products) in similar signaling pathways [129]. Therefore, patients carrying KRas mutations have shown partial response to the EGFR targeted treatments and thus are considered as negative prognostic biomarkers [128]. Moreover, the signal intensity of the wild-type peptide (figure II.1A) includes the NRas and HRas isoforms. If the genomic profile of a cell line is unknown, a clear differentiation between the isoforms cannot be obtained only from the proteomics profiles [130]. To be exact, it is not clearly defined if the detected signal of KRas only results from the KRas mutation or if the other two isoforms contributed to its intensity. On the contrary, precise differentiation of the isoforms at genomic level was not done due to the predominant DNA coding sequence of KRas resulting in poor protein translation and subsequent activation of different genetic events if mutation occurs [131]. Moreover, one proteomic study using knockdown SW48 colon adenocarcinoma cell lines showed that the abundance pattern of the isoforms was KRas>NRas>HRas [132]. The authors of these study used similar approach as ours, combining IP with MS working in selected reaction monitoring (SRM) mode. They found this approach lacking in sensitivity due to the detection limits (< 25 fmol/mg or 6000 cells) that might be insufficient for detection of mutations in clinical samples. However, our PRM analysis showed KRas>NRas>HRas pattern only in the A549 and H3255 cells, whereas the H1975 and HCC827 cell lines were expressing the NRas isoform the most (46% and 54% respectively). Furthermore, our IP-PRM targeted analysis were done on around 4000 cells.

On the other hand, the EGFR mutations and their wild-type signature peptides were generated by GluC endopeptidase, which cleaves after glutamic (E) and aspartic (D) acids thus producing peptides suitable for targeted PRM analysis. Although this protease is prone to missed-cleavages (as described in [95]), the SIL peptides were synthetized to aid the identification and quantification of the mutation status in each of the five cancer cells. As expected, the deletion mutation was identified in the HCC827 cell line at 92% mutation rate, whereas all the other cells were described as homozygotes due to the 100% expression of the KVKIPVAIKE wild-type peptide. Furthermore, the L858R mutation was quantified in the

H1975 and H3255 cells (55% and 92%, respectively), confirming the high sensitivity and selectivity of the IP-PRM approach. As mentioned before, mutations in oncogenic EGFR occur in heterozygous settings, where the mutant to wild-type allele ratio may be imbalanced [122]. The estimated EGFR mutation status in H3255 cells confirmed the domination of the mutant allele over the wild-type one. This allele-specific suppression is often found in NSCLC mutated cancer cell lines, where the mutant allele is always overexpressed [133] and is more associated with the EGFR deletion mutation [134]. However, the mutant allele-specific imbalance was found to provide only advantage to the tumor cells, without impact on the targeted therapy [4]. The developed methodology was found suitable for assessment of the EGFR status, due to the ability to process multiple different samples at once. Likewise, information regarding the performance of different drugs, targeting multiple primary and secondary mutations as well as characterization of PTMs present in plasma, cell or tissue samples can be obtained with high accuracy and throughput [119].

Lung cancer as a heterogeneous disease requires comprehensive and in-depth profiling for personalized targeted treatments. The precise assessment of the mutation status of each patient prior therapy selection holds great value. This is especially important in cases when mutated allele is highly activated with the respect to the wild-type demonstrating the therapeutic relevance due to the prediction of the response to the targeting inhibitors [135]. Moreover, the assessment of the accurate EGFR mutation grade (such as occurrence and rate of primary and secondary mutations, metastasis, smoking history, tumor size *etc.*) has to be considered in patients with advanced stage disease who require even third-line therapeutic intervention [136]. As discussed in the published paper on which this chapter is based, the IP-PRM approach demonstrated unambiguous discrimination between the Ras and EGFR targets by estimation of the mutation rates of all mutations and isoforms. Furthermore, due to the modification and optimization presented in Chapter I, *i.e.* the improved cell lysis, the automatized IP step and the adjustment of the chromatographic separation onto a general instrument, this IP-PRM approach offers valuable potential for providing fast results essential for patient stratification in targeted therapies. This quantitative approach could provide information regarding the abundance of mutations and PTMs in normal and cancer cells related to different molecular interactions and signaling pathways with predictive characteristics.

MATERIALS AND METHODS

Chemicals and Reagents

A549, H1975, HCC827, H3255 and A421 cancer cell lines were a gift from the Laboratory of Experimental Haemato-Oncology (LHCE) of the Luxembourg Institute of Health (LIH). MSIA® Disposable Automation research Tips protein A/G (Cat. No. 991PRT15). Monoclonal anti-Pan-Ras antibody, clone RAS 10 (Cat. No. MABS195) was purchased from Merck Millipore. Monoclonal EGFR/ErbB1 antibody (clone 528) (Cat. No. NB110-5846) was purchased from Novus Biologicals. Sequencing grade modified trypsin (Promega) was used for generation of signature peptides for the Ras family isoforms. Endo-Glu-C (Staphylococcus protease V8) was obtained from Worthington for the EGFR representative peptides. Stable isotopically labelled peptides were synthesized by Thermo Fisher Scientific. All other reagents were obtained from Sigma Aldrich.

Sample preparation step

Biological material: A549 lung adenocarcinoma cells were grown in Dulbecco's modified Eagle's medium (DMEM/F12) (Lonza) supplemented with 10% (v/v) fetal bovine serum (FBS) (Life Technologies) and 1% (v/v) penicillin/streptomycin mixture (Lonza). H1975, HCC827 and H3255 lung cancer cells were grown in RPMI-1640 medium (Lonza) supplemented with 10% (v/v) FBS and 1% (v/v) Pen/Strep mix. A431 epithelial cancer cells were grown in DMEM high glucose, pyruvate medium (Life Technologies) supplemented with 10% (v/v) FBS and 1% (v/v) Pen/Strep mix. Cells were incubated at 37°C in 95% humidity atmosphere and 5% CO_2 and grown to complete confluence in T-75 or T-175 flasks. The confluent cells were collected from the flasks by washing with PBS buffer (Life Technologies), followed by incubation with 0.02% (w/v) ethylenediaminetetraacetic acid (Lonza) at 37°C for 15 minutes for cell detachment. The number of cells was estimated using Countess™ automated cell counter (Invitrogen) or manually using Bruker hemocytometer (Sigma). After centrifugation at 300 xg for 10 minutes, the supernatant was aspirated and the cell pellets were used immediately or stored at -80°C.

Immunopurification: Protein extracts were subjected to protein enrichment using 2 µg of mAbs previously loaded on the micro-columns. The immunopurification was performed by repeated aspiration/dispense cycles of the cell extracts through the micro-column for protein

binding to the antibody onto automated liquid handler Versette® working station (Thermo). After capturing the protein, the micro-columns were washed once with 2M ammonium acetate buffer (pH= 8) and 10 mM PBS (pH=7.6) buffer and additionally 7 times with short water wash aspiration/dispense cycles. Ras and EGFR proteins were eluted form the micro-column in 0.4% TFA/40% ACN elution buffer (pH=2.5) and eluates were vacuum dried.

Digestion, IS spiking and desalting: Dried samples were re-suspended in 30 µL of 50 mM ammonium bicarbonate (for Ras proteins) or in 100 mM sodium phosphate buffer (pH=7.8) (EGFR) and reduced with 50 mM DTT for 45 minutes at 50°C and alkylated with 150 mM IAM for 30 minutes in the dark. 0.05 µg trypsin (Ras proteins) or 0.04 µg of GluC protease (EGFR) was added to each sample for overnight digestion at 37°C. The next day, samples were spiked with heavy labelled peptides, desalted onto solid phase extraction Sep-PakC18 cartridges (Waters), vacuum dried and re-suspended in 25 µL of 0.1% formic acid in water for LC-PRM analysis.

LC-MS targeted analysis

LC separation: The chromatographic separations were done on a Dionex Ultimate 3000 RSLC chromatography system, operating in a column switching setup. The mobile phase A consisted of 0.1% formic acid in water and the mobile phase B of 0.1% formic acid in acetonitrile. The loading phase of the samples was composed of 0.05% trifluoroacetic acid and 1% acetonitrile in water. Samples were injected and loaded onto a trap column (75 µm x 2 cm, C18 pepmap 100, 3 µm) at 1 µL/min or 5 µL/min, followed by elution onto an analytical column (75 µm x 15 cm, C18 pepmap 100, 2 µm) with 300 nL/min flow rate. Separation was done by a linear gradient starting from 2% to 90% B in 37 min.

PRM analysis: The PRM analyses were performed on a QExactive Plus (Thermo Scientific) mass spectrometer equipped with an EASY-spray ion source. The PRM method was performed with a quadrupole isolation window of 1 m/z units, an automatic gain control target of 1e6 ions, a maximum fill time of 250ms and an orbitrap resolving power of 70000 at 200 *m/z*. Collision energy was optimized for each precursor. The duration of the scheduled time windows for each pair of endogenous and heavy labelled peptides was set to 2 min.

Data processing: Fragment ion chromatograms were extracted from the MS raw data and processed using the Skyline package software version 3.7.0.11317. Fragment ions were selected according to the accuracy of the mass measurement and the co-elution and corresponding fragment patterns between the endogenous and isotopically labeled standards. For each peptide, the ratios between the sum of the fragments of the endogenous peptides and the labelled ones were calculated.

Chapter III

Comparison with current techniques

BACKGROUND

The majority of observed genomic alterations in cancer are found to be somatically acquired "driver" mutations, characteristic for each cancer type [9]. Identification of this small subset of variations that play an important role in tumor initiation and progression has a strong clinical relevance due to their impact on targeted anticancer therapies and related patient stratification.

In current clinical settings, direct Sanger sequencing is considered the "gold standard" for DNA mutation testing [137-140], although this method lacks in analytical sensitivity for mutation detection in samples containing less than 25% mutated cells [141]. On the other hand, Next Generation Sequencing (NGS), displaying high-throughput, sensitivity, accuracy and multiplex capabilities for deep DNA and RNA sequencing, has still limited use for research due to methodological complexity and the need for validation prior full implementation in clinical practice [142-145]. Other widely used approaches for genomic and transcriptomic mutation analyses are polymerase chain reaction (PCR)-based methodologies. Even though offering quantitative, rapid and more sensitive studies, and the ability to detect mutations with less than 10% frequency, these methods can exhibit primer binding non-specificity and can only monitor one target at a time [146-148]. Alternatively, when sample availability and quality are limited, immunohistochemistry (IHC) analysis can offer high sensitivity and selectivity towards mutations present with less than 10% rate [149, 150], yet limited by the availability of mutation-specific antibodies [151, 152]. Although, there are diverse methods – such as FISH [153], dHPLC [154], ARMS [154, 155] *etc*. – available for screening and targeted analysis with various sensitivities, capabilities and restrictions, they are still restricted for routine clinical use due to their cost, low specificity, antibody availability, complexity, requirements for skilled personnel and long turnaround time [67, 156, 157].

Beside the genetic modifications involved in tumor initiation, progression and pathogenesis, reversible DNA alterations known as epigenetic changes can influence the transcriptional step of the genetic information flow. These changes mainly found as DNA methylations can cause gene activation or silencing, thus resulting in mis-regulated gene function and expression [158]. These changes may have an impact on the RNA and protein expressions as products of the modified genes, hence influencing the effects of the targeted anticancer drugs against the activated oncoproteins that need to be inhibited [159-161]. The genomic

and transcriptomics events do not always correlate with the abundance and expression of proteins, which are ultimately the “driving” force of the cells. Therefore, investigation of genomic events together with protein profiles, cellular signaling pathways and presence of modifications (as mutations and/or post-translational modifications) is beneficial for accurate prognostic and predictive insights into the disease.

Tumor heterogeneity and availability of modified DNA content represent two of the main limitations of mutational analysis [162]. The need to detect and identify all the alterations present within a sample requires significant amounts to be accessible for multiple testing. Hence, effective, multiplexed assays with high-throughput, sensitivity and selectivity are required for analysis of driver mutations. Mass spectrometry-based proteomic analysis can provide information regarding the protein mutations and post-translational modifications related to the disease in a qualitative and quantitative manner by measuring the mass-to-charge (*m/z*) ratio of representative ions. Investigating the disease related changes at protein level results in better sequence coverage, ability to differentiate between isoforms within a single gene and analysis of multiple targets in one batch of analysis, with high sensitivity, selectivity and reproducibility [95, 163-165]. Moreover, protein-based discoveries can contribute to the current genomic knowledge and provide better understanding of the alterations events occurring during the information path from gene to protein [166].

Molecular profiling of a cancer type implicates screening individual genes for the presence of driver mutations to predict the tumor’s response to the targeted drug. For instance, the presence of an activating mutation in the tyrosine kinase domain of epidermal growth factor receptor (EGFR) in non-small cell lung cancer (NSCLC) is associated with sensitivity to the kinase inhibitors erlotinib and gefitinib [167, 168]. These mutations are mainly found as in-frame deletions in exon 19 or as missense substitutions in exon 21, exons that encode the tyrosine kinase domain of the receptor [169]. Estimation of their mutation frequency is important for an accurate prognosis of the disease [170]. Lung cancer diagnosis and classification are currently based on cellular morphology and histological structure [171]. In this manner information regarding the disease origin, tumor type and stage as well as tissue occupancy are obtained as relevant information for patient therapies [172]. This information describing molecular changes and unique signatures is not associated with protein

expression, thus misleading patient stratification and their later outcome by providing false-positive results [173].

Due to poor prognosis, especially for patients in an advanced disease stage, and to obtain better insight into the cancer, we planned to monitor the connection between EGFR genomics, transcriptomics and proteomics events. Four lung adenocarcinoma cell lines, A549 harboring wild-type EGFR (EGFRwt), HCC827 having the deletion ($EGFR^{delE746\text{-}A750}$) mutation and H1975 and H3255 cell lines carrying the point ($EGFR^{L858R}$) mutation, and A431 epidermoid carcinoma cell line with EGFRwt (as control) were used for assessment of the EGFR gene status and protein expression. The presence and mutation rates of EGFR at DNA, RNA and protein levels were evaluated using PCR-based methods, Sanger sequencing and mass spectrometry, techniques with diverse sensitivities towards the detection of mutations. The ability to measure and quantify "driver" mutations at protein level with high sensitivity and selectivity, besides providing the "real picture" of the disease, can provide additional information to genomics results helping to confirm the patient selection for personalized targeted therapies.

RESULTS

EGFR gene copy number variation, DNA content, mRNA and protein expression

The EGFR status in four lung adenocarcinomas and one epidermoid carcinoma cell lines harboring EGFR wild-type, exon 19 deletion and exon 21-point mutation was evaluated at gene, mRNA and protein level as shown in figure III.1. First, the array comparative genomic hybridization (aCGH) analysis on DNA extracted from the cell lines selected for this study showed that EGFR gene was highly amplified in the A431, HCC827 and H3255 cell lines (in decreasing order of copy number (CN) values, albeit out of the linear range of the CGH method) contrarily to A549 cells (CN=3 due to trisomic chromosome 7) and H1975 cells (CN=2.8) (figure III.1A). Then, real-time PCR data, using genomic DNA (gDNA) and primers that bind EGFR exon 19 outside of the deletion sequence, confirmed the increased EGFR DNA amount in HCC827, H3255 and A431 cells when compared to A549 and H1975 cells (presented as (35-Cq) values; 35 being the highest Cq value for DNA to be considered as detected by PCR) (figure III.1B). Further, RT-PCR analysis was performed on complementary DNA (cDNA) from RNA extracts from the selected cell lines.

EGFR mRNA levels were normalized to RNA levels of 4 housekeeping genes (18S rRNA, EEF1a1, GAPDH and Ezrin) and displayed as fold change relative to the A549 cell lines. This analysis indicated increased relative mRNA expression in the three cell lines with higher DNA content. Specifically, the highest relative mRNA expression was observed in the A431 cells, followed by H3255 and HCC827 cells compared to the A549 and H1975 cell lines (figure III.1C). Lastly, investigation of the EGFR protein expression in total cell extracts from the described cells by mass spectrometry-based analysis demonstrated EGFR protein overexpression in A431, H3255 and HCC827 cells compared to the A549 and H1975 cell lines (figure III.1D). The EGFR protein overexpression in these three cell lines was confirmed by western blot analysis of cell extracts targeting the total EGFR protein (figure III.1E). The EGFR gene CNV, mRNA and protein overexpression were highest and consistent in the A431 cells; however, this cell line showed less DNA content than the HCC827 cells. Differences were observed between the mutated HCC827 and H3255 cell lines. Namely, the EGFR gene was more amplified in the HCC827 cell line compared to the H3255 cells, while both cell lines displayed similar EGFR mRNA and protein expression. The mRNA expression was related to the protein expression in all cell lines. These results demonstrate that the gene CNV can only partially explain the differences observed at mRNA and protein expression level in relation to the DNA content.

EGFR exon 19 deletion mutation rate

One of the most activating and frequent EGFR mutations in lung cancer occurs as an in-frame deletion in exon 19, the exon that encodes part of the tyrosine kinase domain. The occurrence and mutation rate (% of present mutation) of the c.2236_2250del15 deletion mutation in the selected cell lines were assessed at DNA level by digital PCR, the relative expression of this allele at RNA level was evaluated by RT-PCR and as EGFR p.E746-A750delELREA mutation at protein level by targeted MS analysis. Digital PCR analysis was initially performed on DNA extracts from the selected cell lines. The quantitative analysis using specific TaqMan dye-labeled probes for detection of the specified deletion indicated occurrence of the activating mutation only in the DNA of HCC827 cell line with a 97.2% mutation rate (figure III.2A). As a confirmation of these results, real-time PCR analysis was performed using primers either EGFR exon 19 wild-type, or EGFR exon 19del.

A

Cell line	A549	H1975	HCC827	H3255	A431
EGF copy number	3	2,8	Highly amplified* (+++)	Highly amplified* (++)	Highly amplified* (++++)

* EGFR copy number values out of the linear range of the CGH method

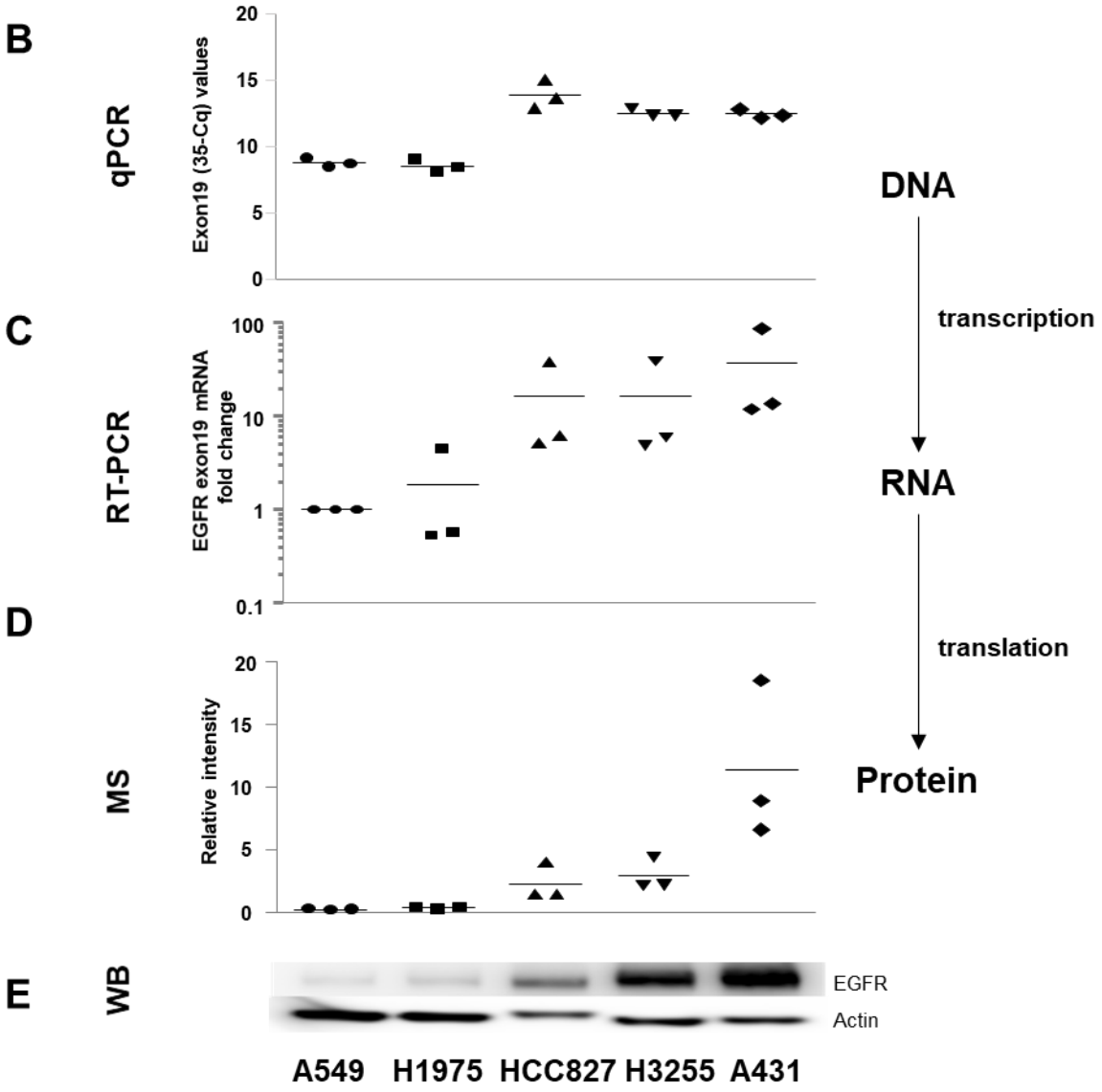


Figure III.1. Estimation of the EGFR gene copy number variation, DNA content, mRNA and protein expression in five cancer cell lines. (A) aCGH copy number variation indicating the EGFR gene amplification in A431, H3255 and HCC827 cells compared to A549 and H1975. (B) Real-time PCR Cq values representing higher EGFR gene expression (exon 19) in A431, H3255 and HCC827 cells. Scatterplots present (35-Cq) values from n=3 complete independent biological replicates. (C) RT-PCR ΔCq values presenting higher mRNA expression of EGFR exon 19 in A431, H3255 and HCC827 cells. Scatterplots present fold change normalization to A549 from n=3 complete independent biological replicates. (D) MS analysis of EGFR protein expression. Scatterplots show the endogenous/heavy peptide area ratio of the GEPREFVE, a surrogate peptide of EGFR, presented as mean ± standard deviation calculated from triplicate biological replicates. (E) Western blot analysis of total protein extracts showing the EGFR and Actin expression, using antibody towards the total EGFR. MS and WB analyses show higher EGFR expression in A431, H3255 and HCC827 cells compared to H1975 and A549.

The mutated allele was identified in the HCC827 cell line as shown by the high (35-Cq) values, whereas EGFR exon19del could not be detected in the other cell lines (Cq > 35) (figure III.2B, right graph). In accordance with the dPCR results that imply the presence of the unmutated EGFR exon19 allele at approximately 3% of total alleles, EGFR exon 19 wild-type DNA was also detected by PCR in the HCC827 cell line (figure III.2B, left graph). Unsurprisingly, high levels of EGFR exon19 wild-type were detected in H3255 and A431 cells due to the EGFR gene amplification described before. Further, to investigate the relative EGFR mRNA expression in these cell lines, RT-PCR analysis was carried out using a SYBR Green detection approach with primers to distinguish between the exon 19 deletion (figure III.3A, right graph) and wild-type EGFR (figure III.3A, left graph). Indeed, all the cell lines but HCC827 displayed RT-PCR Cq values near or above 35, whereas mean Cq values of housekeeping genes (HKG) did not vary considerably (total mean of HKG Cq values ± 95% confidence interval=19.27 ± 0.22, 19.05 ± 0.78, 19.39 ± 0.96, 19.23 ± 1.20 and 19.31 ± 0.69 for A549, H1975, HCC827, H3255 and A431, respectively). As for EGFR exon 19 wild-type mRNA, it was expressed in all the cell lines at variable intensities as shown by the fold change values in figure III.3A (left graph).

A dPCR

Mutation \ Cell line	A549	H1975	HCC827	H3255	A431
c.2236_2250del15	Negative	Negative	97.2%	Negative	Negative

B RT-PCR

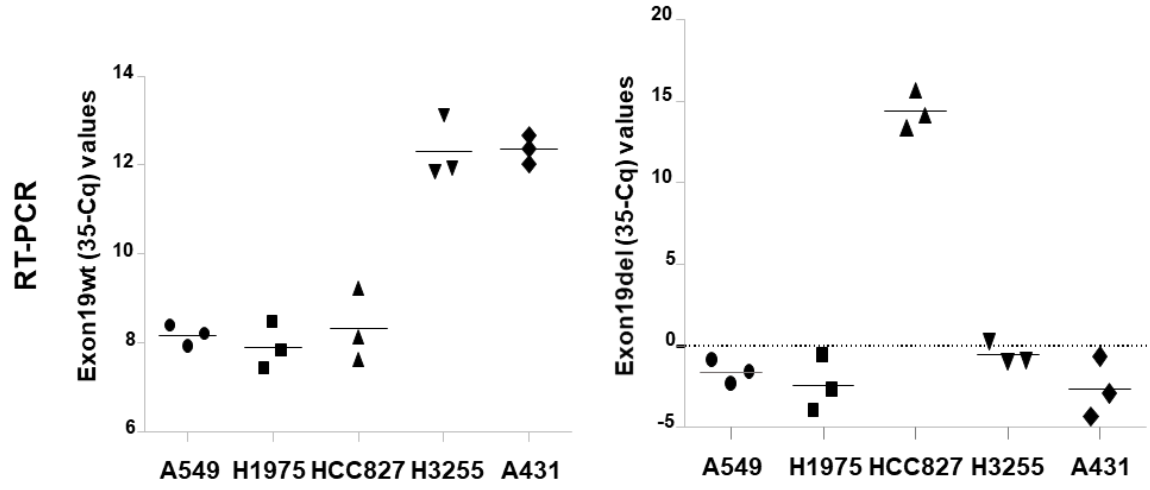


Figure III.2. Estimation of the DNA exon 19 *del*E746-A750 mutation rate in the cancer cell lines. (A) dPCR analysis of DNA extracts from the cells reporting detection of the exon 19 mutated allele only in the HCC827 cell line. (B) Cq values obtained by RT-PCR analysis using gDNA presenting the EGFR exon 19 wild-type (left graph) and the deletion mutation (right graph). Scatterplots present (35-Cq) values from n=3 complete independent biological replicates. The mutation was detected only in the HCC827 cells.

At last, the *del*E746-A750 mutation rate at protein level was evaluated by targeted mass spectrometry analysis on digested protein extracts. The quantitative MS analysis identified the representative peptide for the deletion mutation only in the HCC827 cells with a 94% mutation rate (figure III.3B, right graph), identifying this cell line as a heterozygote due to the identification of the wild-type representative peptide with 6% mutation rate (figure III.3B, left graph). The A431 and H3255 cells were identified as 100% wild-type, whereas no signals were measured in the A549 and H1975 cells. The estimated exon 19 mutation occurrence and rate in HCC827 cells were consistent between the analyses, despite the observed differences among the EGFR gene CNV and mRNA and protein expression in this cell line.

EGFR exon 21-point mutation rate

The second most frequent EGFR sensitizing mutation in NSCLC occurs in exon 21 of the tyrosine kinase domain of the biomolecule. This activating mutation is a one base pair T>G substitution in the DNA sequence at position 2573 resulting in one amino acid substitution (leucine-to-arginine) at position 858. Identification and quantification of this EGFR mutation was performed at gene and transcriptome levels by digital PRC technology and at protein level using targeted mass spectrometry analysis in extracts from the selected cancer cells. First, the gDNA extracts from the cell lines were analyzed by dPCR where specific dye-labeled probes for detection of the c.2573T>G substitution mutation were used. This approach allowed the detection of the mutation in the H1975 and H3255 cell lines with mutation rates of 78.3% and 93.6%, respectively (figure III.4A). In parallel, Sanger sequencing was carried out as a verification of the T>G substitution in the DNA extracts from the described cell lines. As illustrated by the electropherograms presented in figure III.4B, the one base pair substitution was identified in the corresponding two cell lines; the H1975 cell line was classified as a heterozygote as also the wild-type allele was detected, whereas the H3255 cells were characterized as mutant.

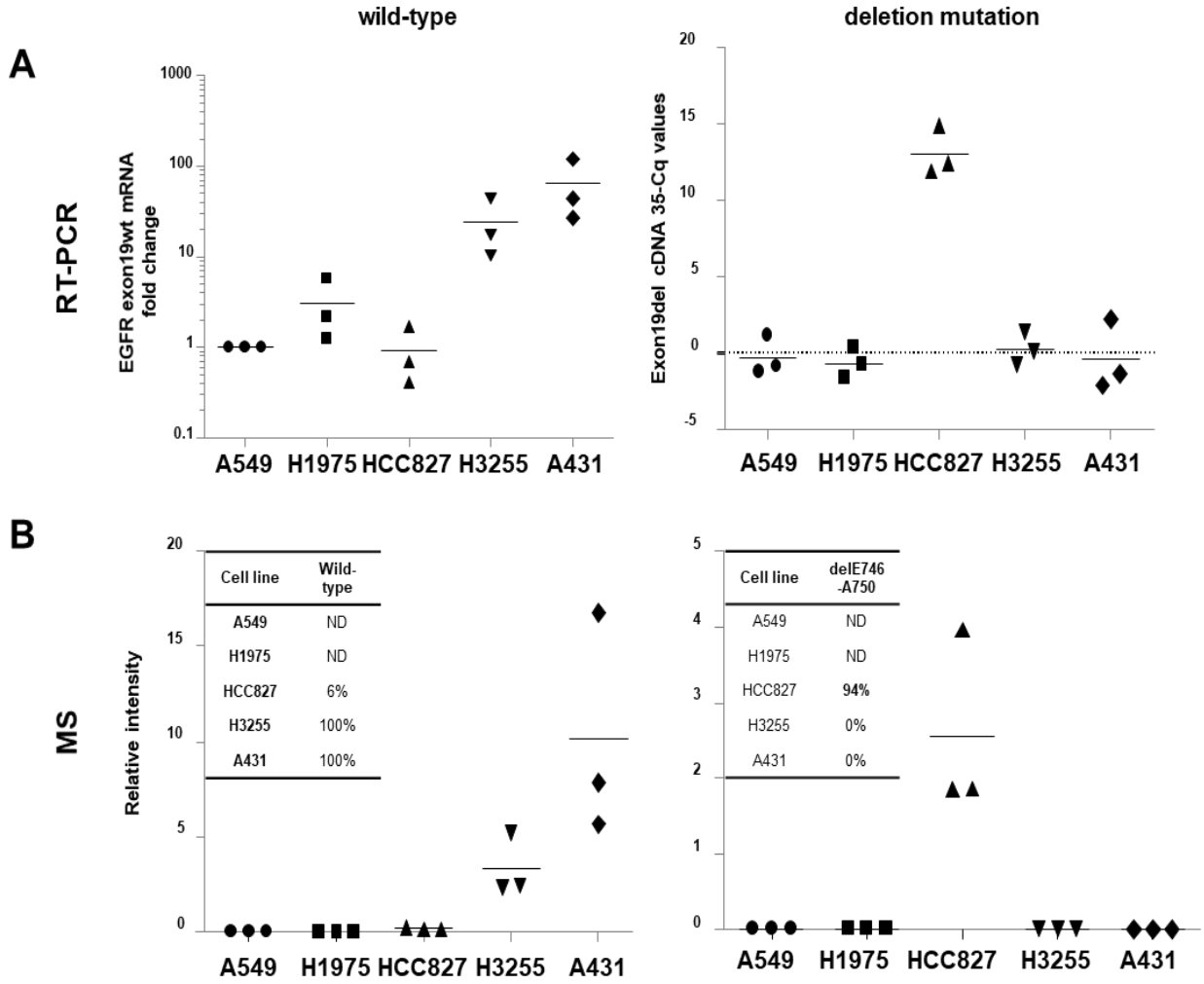


Figure III.3. Estimation of the mRNA and protein exon 19 deletion mutation rate in the studied cells. (A) EGFR mRNA expressions (normalized to HKG) in the cell lines measured by RT-PCR analysis presenting the EGFR exon 19 wild-type (left graph) and the deletion mutation (right graph) calculated as fold change (for wild-type) and (35-Cq) (for mutation) values from n=3 complete independent biological replicates. (B) Estimation of the EGFR *del*E746-A750 mutation rate in the cell lines by targeted MS (protein level). Scatterplots present as mean ± SD variations calculated from n=3 complete independent biological replicates. The mutation was detected only in HCC827 cells.

Second, the dPCR was performed using cDNA from the above-mentioned cell lines for the quantification of the EGFR L858R mutation at the transcriptomic level. The EGFR L858R cDNA was detected only in H1975 and H3255 cells (Figure III.5). The percentage of the mutated allele at mRNA level in these cell lines was 81.2% and 93.1%, respectively (figure III.5A). Third, MS-based analysis was performed measuring the generated representative peptides for the L858R mutation and the corresponding wild-type as surrogates for the EGFR protein.

A

dPCR

Mutation \ Cell line	A549	H1975	HCC827	H3255	A431
c.2573T>G	Negative	**78.3%**	Negative	**93.6%**	Negative

B

Sanger

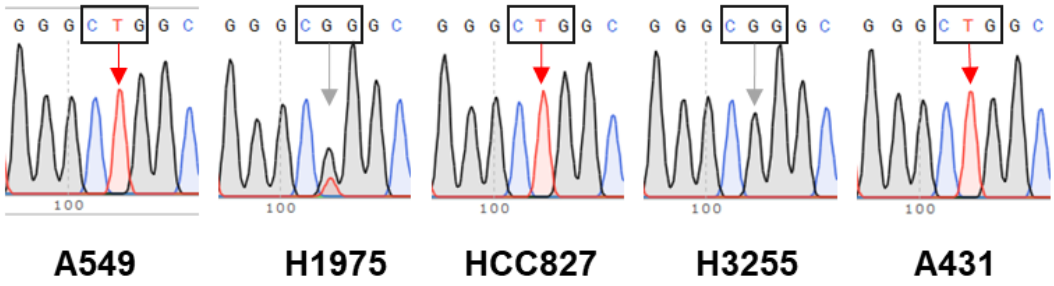


Figure III.4. Estimation of the exon 21 L858R mutation rate in the studied cells. (A) dPCR analysis of DNA extracts from the cells reporting detection of the exon 21 mutated allele only in H1975 and H3255 cells. (B) Sanger sequencing analysis of the T>G substitution mutation in the cells. The mutation along with the wild-type was detected in the H1975 cells, where H3255 was characterized as mutant.

The quantitative targeted proteomics analysis identified the representative peptide for the L858R mutation only in the H3255 cell line with mutation occurrences of 88%, respectively (figure III.5B, right graph), while the mutation was not detected in H1975 cells, probably due to the low abundance of the protein. The corresponding wild-type representative peptide was detected only in A431, H3255 and HCC827, cell lines overexpressing EGFR (figure III.5B, left graph). The genomics, transcriptomics and proteomics exon 21 L858R mutation analyses demonstrated concordance between the obtained mutation rates in the H3255 cells, showing the higher sensitivity of the dPCR technique over the standard Sanger sequencing method. On the contrary, mutation was not detected at protein level in the H1975 cells compared to the mutation rates at DNA and RNA (78% and 81%, respectively) levels in the same cell line.

A

dPCR

Mutation \ Cell line	A549	H1975	HCC827	H3255	A431
c.2573T>G	Negative	**81.2%**	Negative	**93.1%**	Negative

B

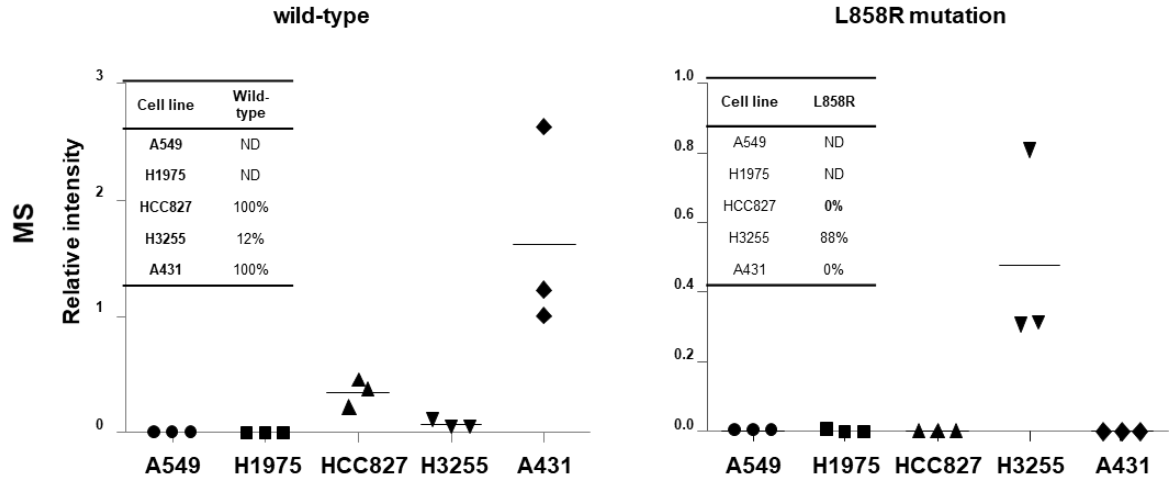


Figure III.5. Estimation of the mRNA and protein exon 21 L858R mutation rate in the studied cells. (A) EGFR mRNA mutation rate obtained by dPCR analysis presenting the EGFR T>G mutation rate in H1975 and H3255 cells. (B) Estimation of the L858R point mutation rate in the cell lines by targeted MS (protein). Bars present as mean ± SD variations calculated from n=3 complete independent biological replicates. Mutation was detected only in H1975 and H3255 cells.

DISCUSSION

Epidermal growth factor receptor is considered as a predictive biomarker for non-small cell lung cancer patients undergoing tyrosine kinase inhibitor (TKI) therapy. The presence of activating mutations in the TK domain of EGFR, mostly in exon 19 as a deletion mutation and in exon 21 as a missense mutation, predicts the sensitivity to TKIs and the prognosis in advanced disease stages. Although these features are helpful, the selection of patients who will benefit from an individualized therapy is still a major problem. Currently, patients are selected according to their individual genomic profiles, whereas predictions of the tumor

response are estimated depending on the efficacy of the drug targeting a specific protein [174].

To improve the patient stratification, instead of searching for the presence of individual mutations, the analysis of multiple parameters with prognostic or predictive value should be performed, including estimation of gene copy number variations and DNA amplifications associated to protein expression. It has been reported that mutations initially occur during tumor development while DNA amplifications and copy number variations are acquired later in the tumor progression [175, 176]. Furthermore, despite the presence of EGFR mutations, increased gene copy number has served as a predictor for sensitivity to TKIs in advanced disease, while showing poor prognostic value. Correlation between the occurrence of EGFR mutations followed by gene amplification later in the pathogenesis has been confirmed [176, 177], and can be partially foreseen by protein abundance. Furthermore, the existence of EGFR mutations might also indicate presence of mutant allele-specific imbalance (MASI) and both are associated with mutant allele transcription and gene activity [122]. In oncogenes, MASI besides to the mutations is as well related to the gene copy number and the tumor heterozygosity, *i.e.,* mutant versus wild-type expression ratio. All these parameters acting together have greater biological and clinical relevance in the tumor development and progression then any individual alteration [4]. When using oncoproteins as drug targets, their expression can be affected by the treatment resulting in downregulation of the protein or acquired resistance to the TKIs [178, 179]. Our results demonstrated that samples with increased EGFR protein expression also showed increased gene copy number and amplification, confirming the relationship among these predictive parameters; the connection between the presence of a mutation and increased amplification was not completely supported since the cell line with the highest EGFR copy number was overexpressing the EGFRwt. On the other hand, protein expression levels were associated with the mRNA expressions in all the cell lines but were not found proportional to DNA abundances. Namely, we observed decreased DNA content in A431 cells harboring EGFRwt while they displayed the highest CNV among the cell lines. The aCGH analysis of these cells showed the presence of an EGFR deletion subpopulation possibly due to unbalanced chromosomal rearrangements which caused reduced DNA content in these cell lines (figure III.6) [180].

The assessment of these predictive parameters in complex tumorous specimen with limited mutant material available is a challenge. Many techniques have been employed over the years for EGFR mutation detection; however, the obtained results are inconsistent when compared between each-other and with the standard direct sequencing method [67].

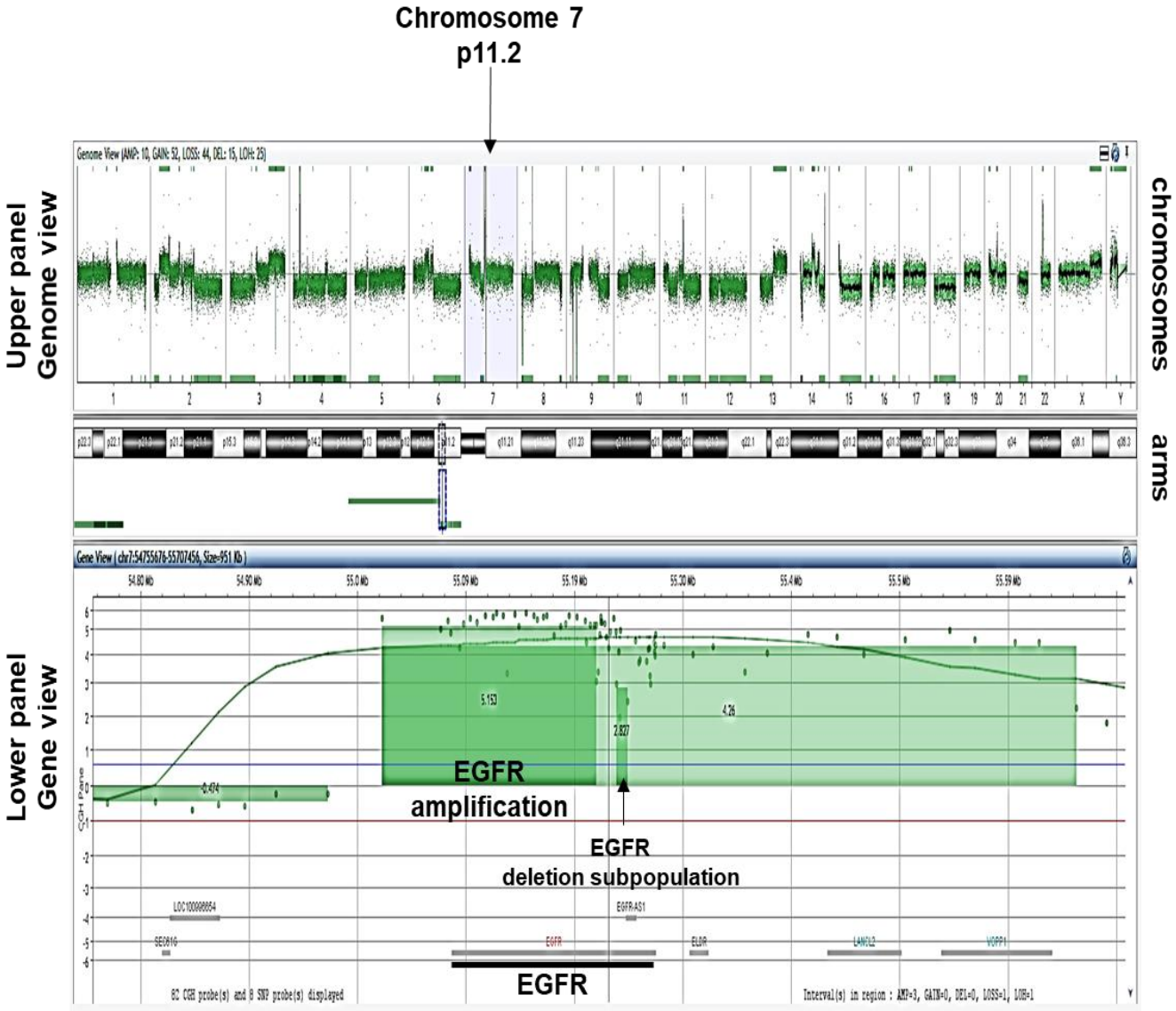


Figure III.6. aCGH profiles of total EGFR in the A431 cell line representing the DNA copy number variation (CNV) patterns for chromosome 7. The upper panel presents the genome view, where chromosome 7 is highlighted with the short p11.2 arm. The lower panel presents the EGFR's gene view (CGH pane), where the region between 54.8 and 55.59 Mb is displayed. EGFR molecular location (55.02 – 55.21 Mb) showed high gene amplification, with additional deletion subpopulation.

Limit of detection and biological sensitivity are the most important features when choosing a method for heterogenic mutation analyses. Additionally, short turnaround time, low DNA input, accuracy and reproducibility are desirable requirements for clinical application. The commonly used methodology for genetic mutation analysis is Sanger direct sequencing of PCR products. The inability to distinguish between cancer subtypes (mutation vs. wild-type) for a low mutation rate (below 25%) is the main drawback of the sequencing method despite its clinical importance. This limitation was confirmed by our results, as the EGFR wild-type allele was not detected in H3255 cells, since it was expressed with <10%. We identified this cell line as heterozygote with mutation vs. wild-type expression 94% vs. 6%, whereas usually this ratio is around 50:50. The real-time polymerase chain reaction-based technique is another widely used approach due to its speed, sensitivity and low sample input features compared to Sanger sequencing. To date, RT-PCR represents the method of choice for detection and quantification of nucleic acids, used alone or coupled with other methods like next-generation sequencing (NGS) [139]. Nonetheless, PCR quantification is mainly based on the measurements of housekeeping genes as normalizers serving as references under the assumption that their behavior is unchanged and unaffected during the measurements [181]. Despite the imprecisions that can occur during DNA/RNA extraction steps primer binding non-specificity is an additional critical point influencing the quantitative performance of the PCR technique. Therefore, to overcome this limitation, the digital PCR method was applied for EGFR mutation detection and absolute quantification with over 95% sensitivity and no prior information requirements regarding the mutation sequence [157]. Applying this approach on the DNA and RNA extracts from the selected cell lines, the exon 19 DNA mutated allele was quantified in HCC827 cells, and the L858R mutated DNA and mRNA in H1975 and H3255 cell lines, overcoming the PCR primer binding issues and confirming the better sensitivity and selectivity of the dPCR method. However, the efficiency of the reverse transcription step itself – including enzymes and suitable reagents – can generate variability in the sample quantification [181, 182].

Regarding the evaluation of the EGFR mutation status in NSCLC, the method of choice should be able to unambiguously identify the cancer subtypes, *i.e.,* to assess the mutation and the wild-type counterpart. Many studies pointed out that patients who harbor deletion or point mutations responded better towards EGFR TKIs compared to those having EGFRwt [183]. However, investigation of gene changes can only predict cancer cell events, while examination of the protein and their post-translational modifications, provides an

accurate view of cell events. Protein analyses allow higher sequence coverage, better detection of protein variations at amino acid level and the ability to distinguish between diverse isoforms within a gene. Protein analysis – including protein profiles, protein-protein interactions and cellular pathways – in addition to investigation of genetic expression, can provide better predictive insights of the disease, leading to faster diagnosis and accurate treatment choices [184]. Mass spectrometry-based methods have been applied in clinical areas mainly for therapeutic drug monitoring and toxicological analyses [185]. The major advantage of this approach is the ability to directly identify and quantify specific protein isoforms in a high-throughput fashion, where mutated and wild-type protein sequences can be monitored simultaneously in a relatively short turnaround time. In addition to the measurement of protein expression changes, MS allows quantitative examination of post-translational modifications related to signaling pathways and networks. Moreover, the MS platform overcomes the limitations of the immunoassay-based protein analyses (as ELISA, WB, protein arrays and IHC) such as antibody specificity and protein dynamic range [186]. For example, in our study, EGFR mutation and wild-type expressions and rates were altogether identified and quantified by a targeted MS approach, while WB was only used for confirmation of the protein presence.

Concerning the above mentioned molecular profiling characteristics, EGFR exon 19 mutational analyses pointed out the importance of the examination of multiple parameters with predictive features. The high copy number and DNA content were not associated to the mRNA and protein expressions, although the mutation rate was consistent at all levels. Additionally, the copy number gain related to the mutant allele-specific imbalance are specially associated to the deletion mutation [122, 187], these two factors did not demonstrate increased mutant allele transcription and gene activity. For instance, if a patient is selected for therapy according to these genomic profiles, he/she might not benefit completely from the drug dose and the toxic side effects might increase [22]. The ability to monitor simultaneously the mutation vs. wild-type protein expressions and rates provides a better insight into the cancer state. On the other hand, exon 21 mutational profiling demonstrated low MS sensitivity toward the A549 and H1975 cell lines due to the low protein abundance in these two cell lines. This limitation can be overcome using protein enrichment prior targeted MS analysis (as described in chapter II of the thesis), where the L858R mutation was identified in the H1975 cells with 55% mutation rate (figure II.3). If we compare these IP-PRM results with the results obtained by dPCR, the differences found in H1975

cells specified the significance of the investigation at protein level. Namely, a decreased L858R mutation rate (81.2% reduced to 55%) was detected in the IP-PRM targeted analysis compared to the dPCR results for these cells, indicating mis-regulation of the translation events leading to a low correlation between the mRNA and protein expressions. Moreover, the allelic suppression of the wild-type allele in the H1975 cell line was also not consistent with the protein expression. Besides the assay conditions under which RNA translation might be reduced, low protein abundance can occur due to the decreased protein half-life, protein degradation and/or presence of post-translational modifications [188].

In conclusion, DNA testing can detect the presence of mutations, but protein profiling will provide precise insight of the disease. Genomic analyses are good indicators to identify potential candidates for follow-up investigations, whereas protein profiles will define the outcomes and benefits of a targeted therapy. Moreover, having the power to deliver information about the proteins of interest and their structural modifications without any prior knowledge, mass spectrometry demonstrated itself as the most sensitive choice for protein analysis.

MATERIAL AND METHODS

Biological material: The five cancer cell lines harboring different EGFR mutations described in Chapter I were used as biological material for the genomics, transcriptomics and proteomics analyses.

Cell lysis and protein extraction: The five cancer cell lines were lysed as described in Chapter I and the EGFR protein was extracted for subsequent LC-PRM analysis.

EGFR Genomic analysis: DNA were extracted from biological samples using DNAeasy® Blood & Tissue Kit (Qiagen Cat# 69504) according to the manufacturer's instructions. DNA quantities were determined using a Nanodrop Spectrophotometer ND-1000 and a Qubit™ 2.0 Fluorometer with the Qubit™ dsDNA BR Assay Kit (Thermofisher Cat# Q32850). For the exon 19 DNA analysis the primers forward 5'-CGCTATCAAGGAATTAAGAGAAGC-3' and reverse 5'-CCACACAGCAAAGCAGAAACT-3'were used for the wild-type and the primers forward 5'-AATTCCCGTCGCTATCAAAAC-3' and reverse 5'-CACACAGCAAAGCAGAAACTCA-3' were used for the deletion. To check the T>G

mutation, EGFR was amplified by PCR using the primers forward 5'-TGATCTGTCCCTCACAGCAG-3' and reverse 5'-AGAGAAACCGAGCCAGTGAA-3'. The amplicons were purified using the MinElute PCR purification kit (Qiagen Cat# 28004) and sent for sequencing to LGC genomics in Berlin (http://www.lgcgroup.com/our-science/genomics-solutions/#.WbZWgMZLeUk). The same forward primer 5'-TGATCTGTCCCTCACAGCAG-3' was used for DNA sequencing. Mutation status for Exon19 deletion and exon 21 mutation were assessed by digital PCR on a QuantStudio 3D Digital PCR System using EGFR Digital PCR Mutation Detection Assays from Thermofisher (Assay ID# Hs000000027_rm for p.E746_A750delELREA and HS000000026_rm for p.L858R) according to manufacturer's instruction. EGFR copy number was evaluated using SurePrint G3 Human Cancer CGH+SNP Microarray Kit 4×180K (Agilent Technologies, ID G4869A) Protocol Version 7.5, June 2016. The data have been analyzed with the software "Agilent CytoGenomics" version 4.0.3.12.

EGFR transcriptomic analysis: Total RNA were extracted from biological samples using RNeasy® Mini Kit (Qiagen Cat# 74104) according to the manufacturer's instructions with addition of the optional DNase digestion step. RNA quantities were determined using the Nanodrop Spectrophotometer ND-1000. RNA integrity was asses by RNA 6000 NanoChips with Agilent 2100 BioAnalyzer. All RNA samples had RIN number > 8.5. 1.5 µg of total RNA were reverse transcribed into cDNA using an oligo(dT)-Random Primer Mix and the superscript III reverse transcriptase (Thermofisher Cat# 18080085) according to manufacturer's instructions. 1/160 volume of the reverse transcription were used to assess real-time quantitative PCR experiment. Exon 19, exon 19 WT and deleted isoform were quantified using a SYBR Green detection (Thermofisher Cat# 4385612) with primers Exon 19 WT forward 5'-CGCTATCAAGGAATTAAGAGAAGC-3', Exon 19 del forward 5'-AATTCCCGTCGCTATCAAAAC-3', and the same reverse primer was used for the two isoforms 5'-GCCATCACGTAGGCTTCATC-3'. To compared with the global amount of EGFR exon 19 couple of primers outside the exon 19 mutation were used. The sequences of the primers used are forward 5'-AGAAAGTTAAAATTCCCGTCGCTAT-3' and reverse 5'-ACGCTGGCCATCACGTAG-3'. Four Housekeeping genes 18S (forwards 5'-TCGAGGCCCTGTAATTGGAA-3' and reverse 5'-GCTGCTGGCACCAGACTTG-3'), EEF1a (forward 5-TTGTCGTCATTGGACACGTAG-3' and reverse 5'-TGCCACCGCATTTATAGATCAG-3'), GAPDH (forward 5'-CATGAGAAGTATGACAACAGCCT-3' and reverse 5'-AGTCCTTCCACGATACCAAAGT-

3') and EZRIN (forward 5'-TGCCCCACGTCTGAGAATC-3' and reverse 5'-CGGCGCATATACAACTCATGG-3') were used to assess relative quantification of EGFR among cell lines using the $2^{-\Delta\Delta Cq}$ method. Exon 21 reverse transcripts by digital PCR as for the genomic analysis, using 900 nmol of reverse and forwards primers and 250 nmol wild-type and mutant probe. The data have been analyzed with the software Thermofisher cloud software.

Western blot analysis: Concentrations of the protein extracts were determined by Qubit™ 2.0 Fluorometer with the Qubit™ Protein Assay Kit (Thermofisher Scientific Inc.). 20 µg of total protein extracts were used for sodium dodecyl sulfate polyacrylamide gel electrophoresis separation followed by their membrane transfer using iBlot Dry Blotting System according to the manufacturer instructions (Invitrogen, Life Technologies). The blots were then incubated with EGF Receptor (D38B1) XP Rabbit mAb (#4267, Cell Signaling Technology) and Peroxidase AffiniPure Goat Anti-Rabbit IgG (H+L) pAb for 4 hours at room temperature. EGFR bands were detected by chemiluminescence substrate SuperSignal™ West Pico PLUS (Thermo Pierce) and Image Quant LAS 4000 system (GE Healthcare, United Kingdom).

Mass spectrometry analysis: Protein extracts were precipitated with ice-cold methanol for two hours. After centrifugation at 20 000 xg at 4°C for 10 minutes, supernatant was discarded, and cell pellets were re-suspended in 10 M urea/100 mM phosphate buffer (pH=7.8), reduced with 70 mM DTT for 45 min at 37°C and alkylated with 220 mM IAM for 30 minutes in dark. 0.5 µg of endopeptidase GluC was added to each sample for overnight digestion at 37°C. Next day samples were spiked with heavy labelled peptides, desalted onto solid phase extraction Sep-PakC18 96 plate format cartridges (Waters), vacuum dried and re-suspended in 100 µL of 0.1% formic acid in water for LC-PRM analysis. The chromatographic separations were performed on a Dionex Ultimate 3000 RSLC chromatography system, operating in a column switching setup. The mobile phases A is consistent of 0.1% formic acid in water and the mobile phase B is consistent of 0.1% formic acid in acetonitrile. The loading phase of the samples was consistent of 0.05% trifluoroacetic acid and 1% acetonitrile in water. Samples were injected and loaded onto a trap column (100 µm x 2 cm, C18 pepmap 100, 3 µm) at 1 µL/min, followed by elution onto an analytical column (75 µm x 15 cm, C18 pepmap 100, 2 µm) with 300 nL/min flow rate. Separation was done by a linear gradient starting from 2% to 90% B in 89 min. The PRM analysis were

performed on a QExactive Plus (Thermo Scientific) mass spectrometer equipped with an EASY-spray ion source. The PRM method was performed with a quadrupole isolation window of 1 m/z units, an automatic gain control target of 1e6 ions, maximum fill time of 250ms and an orbitrap resolving power of 70000 at 200 m/z. Collision energies were optimized for each precursor. The duration of the scheduled time windows for each pair of endogenous and heavy labelled peptides were set to 3 min. For data processing, fragment ion chromatograms were extracted from the MS raw data and processed using Skyline package software version 3.7.0.11317. Fragment ions were selected according to the accuracy of the mass measurement and the co-elution and corresponding fragment patterns between the endogenous and isotopically labeled standards. For each peptide, the ratios between the sum of the fragments of the endogenous peptides and the labelled ones were calculated.

Chapter IV

Post-translational modification characterization

BACKGROUND

The presence of EGFR activating somatic mutations (*del*E746-A750 and L858R) enhances the receptor's tyrosine kinase activity and increases the level of autophosphorylation of key tyrosine phosphorylation sites [189, 190]. Both the mutations and the phosphorylation status might have an impact on targeted therapies due to the activation of diverse signaling pathways upon phosphorylation initiation [191]. Therefore, identification and characterization of the activated tyrosine phosphorylation sites can serve as additional information with predictive value for the selection of patients harboring EGFR activating mutations eligible for TKIs therapy.

Characterization of the phosphorylation as a dynamic post-translational modification includes (1) identification of the activated phosphorylation sites, (2) estimation of their expression levels (site occupancy) within a sample or between different samples and (3) assessment of the dynamic changes over time for each identified phosphorylation site. The general MS-based analytical strategy is designed depending on the experimental subject, and involves cell lysis and protein extraction, generation of peptides, phosphopeptide enrichment and mass spectrometry analysis [192]. These qualitative and quantitative phosphoproteomic informations can be acquired in a single experiment using microgram amounts of protein by following the mass (+80 Da) of the phosphopeptide, its abundance and its ionization efficiency. However, there are key analytical challenges affecting the phosphorylation characterization: suppression of low abundant phosphopeptides and decreased sensitivity due to presence of the phosphoryl group (PO_3^{2-}) [193].

In this study, the relationship between the EGFR mutation status and phosphorylation activation of key tyrosine sites was established by targeted PRM analysis, method described in chapter I. The immunopurification of EGFR at protein level allowed conservation of the entire sequence and identification of the phosphorylated sites in addition to its mutational profiling. Furthermore, the expression level of key tyrosine autophosphorylation sites was estimated and the dynamic profiles of the detected phosphotyrosine peptides were monitored in the A431 (EGFRwt) cell line after different stimulation time points. Stoichiometric analyses including EGF stimulation and dephosphorylation step in the workflow have demonstrated the association between the EGFR L858R mutation and the activation of phosphotyrosine sites [194]. Additionally, our phosphorylation targeted

analyses confirmed the capacity of the established platform to perform mutational profiling and post-translational modification characterization of protein purified from cancer samples.

RESULTS

Identification of EGFR tyrosine phosphorylation sites

Phosphorylation on a tyrosine site corresponds to about 0.05-0.1% of the protein phosphorylation events happening in most cells [195]. These phosphorylated sites are present in low abundance compared to the unmodified residues, are randomly localized in different cell types and their sensitivity depends on the stability of the phosphoryl groups attached to the tyrosine amino acid [196]. Therefore, the most critical parameter for tyrosine phosphorylation identification and subsequent quantitative analysis in our work was obtaining enough amount of phosphorylated EGFR protein for analysis.

To overcome the above mentioned limitations and to obtain high sequence coverage, some general requirements were introduced in the workflow. First, the cell harvesting and lysis steps were performed rapidly on ice, using sodium orthovanadate as tyrosine phosphatase inhibitor to protect the phosphorylated sites. Second, the collected samples, if not lysed immediately, were snap-frozen at -80°C to preserve the protein condition for later analysis. Third, EGFR was purified at protein level but the additional phosphopeptide enrichment and fractionation steps were omitted to avoid sample loss. Fourth, trypsin was used for peptide generation due to its high cleavage specificity, accordingly, tryptic peptides containing the tyrosine phosphorylation and the unmodified counterpart were synthetized for PRM identification and quantification analyses.

EGFR contains 20 tyrosine sites prone to autophosphorylation; 10 distributed in the tyrosine kinase domain without major biological significance and 10 localized in the regulatory domain of the protein interacting with various adaptors and signaling proteins [62]. For identification of these tyrosine sites, 20 phosphopeptides were selected to serve as internal standard controls. However, only 11 (10 singly and 1 doubly) phosphorylated peptides were acquired. The remaining peptides were not feasible to be synthetized due to peptide length (> 35 amino acids), amino acid hydrophobicity and possible modification (*e.g.* cysteine carbamidomethylation). The phosphorylated (pY or pTyr) and unmodified (Y or Tyr) peptide sequences with their MS characteristics are presented in table IV.1. Initially, these SIL

peptides were analyzed alone to observe their LC-MS behavior *i.e.,* retention time, ionization efficiency and detection limit, and to establish the most optimal collision energy (nCE) for each peptide's MS fragmentation. An example is shown in figure IV.1, representing the obtained nCE values for the pTyr1197 phosphopeptide, where nCE=20 was chosen as the optimal collision energy for fragmentation of the precursor ion representative of this phosphopeptide. After optimization of the MS parameters, these SIL peptides were used for identification of the phosphorylation sites in EGFR samples purified from the five cancer cell lines previously described. From the 11 phosphotyrosine targets, only 9 singly phosphorylated peptides (tyrosine residues at positions 727, 827, 869, 978, 1069, 1092, 1110, 1172 and 1197) and one doubly phosphorylated peptide containing residues 1069 and 1092 were detected in all the cell lines, excluding the identification of the phosphorylated tyrosine at position 978 in the H3255 ($EGFR^{L858R}$) cell line.

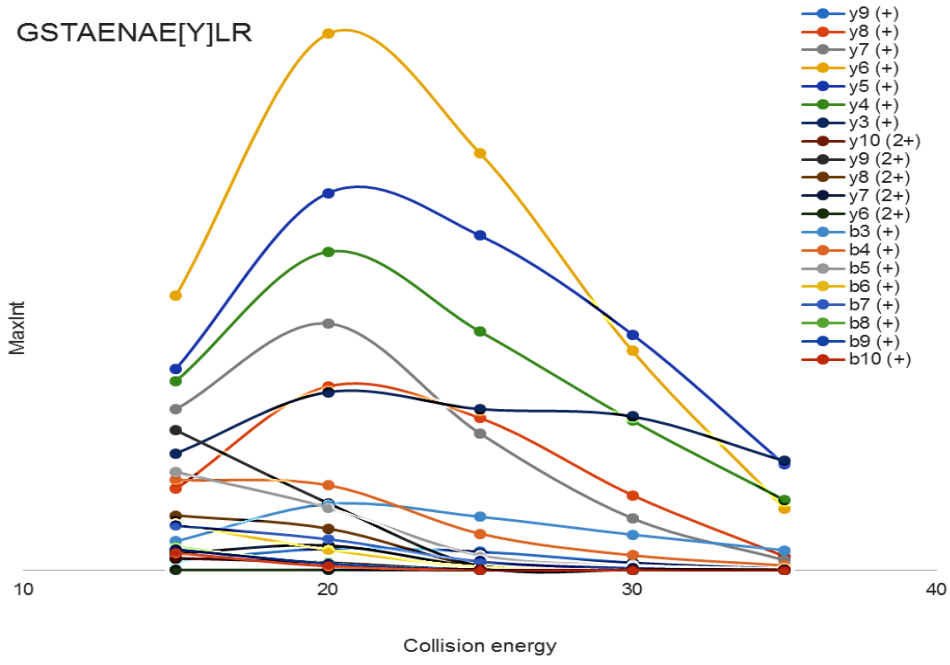


Figure IV.1. Collision energy graph presenting the ion fragment distribution over different nCE values of the phosphorylated peptide GSTAENAEpYLR.

Due to the different expression levels of EGFR among the cells and to compare the phosphorylation expression between the different cell lines, the signal intensities of the identified phosphorylated and unmodified peptides were normalized to the signal intensity of the VAPQSSEFIGA control peptide (peptide nonrelated to any mutation or modification) representing the total EGFR expression in each cell line. The detected and normalized signal intensities of these nine tyrosine phosphopeptides are presented in figure IV.2 together with the intensities of their unmodified counterparts. The different tyrosine phosphorylated sites showed different expression patterns in the cell lines, whereas similar expression patterns of the unmodified peptides were observed in all the cells. Namely, expressions of the pTyr727 and pTyr1110 were similar among all the cells. Phosphorylated tyrosine at position 827 was less expressed in A431 (EGFRwt) and pTyr869 and pTyr1069 in H3255 ($EGFR^{L858R}$) compared to the other cells. The phosphotyrosine at position 978 was not identified in H3255 cell line, while pTyr1092, pTyr1172 and pTyr1197 autophosphorylated sites in the regulatory domain of EGFR were preferentially expressed in these cells. Also, phosphorylated and unmodified peptides of pTyr1172 and pTyr1197 sites showed similar expression patterns in all the cells. The doubly phosphorylated peptide (pTyr1069/1092) was more expressed in A549 (EGFRwt), H1975 ($EGFR^{L858R}$) and HCC827 ($EGFR^{del E746\text{-}A750}$) cells compared to H3255 ($EGFR^{L858R}$) and A431 (EGFRwt) cell lines. This PRM phosphopeptide analysis demonstrated the relation of the pTyr1172 and pTyr1197 autophosphorylation sites to the EGFR L858R mutation.

Table IV.1. EGFR tyrosine phosphorylation sites.

Phosphorylation site	Sequence	Charge	*m/z*	Collision Energy
pY-727	VLGSGAFGTV[Y]K*	2	639.8101	20
Y-727	VLGSGAFGTVYK	2	599.827	15
pY-827	GMN[Y]LEDR	2	539.2072	20
Y-827	GMNYLEDR	2	499.224	15
pY-869_miss**	E[Y]HAEGGKVPIK	3	469.8938	20
Y-869_miss	EYHAEGGKVPIK	3	443.2383	20
pY-978	[Y]LVIQGDER	2	586.771	20
Y-978	YLVIQGDER	2	546.7878	15
pY-998	MHLPSPTDSNF[Y]R	3	548.9007	15
Y-998	MHLPSPTDSNFYR	3	522.2453	20
pY-1016	ALMDEEDMDDVVDADE[Y]LIPQQGFFSSPSTSR	3	1229.851	15
Y-1016	ALMDEEDMDDVVDADEYLIPQQGFFSSPSTSR	3	1203.196	15
pY-1069+1092***	[Y]SSDPTGALTEDSIDDTFLPVPE[Y]INQSVPK	3	1186.855	15
Y-1069+1092	YSSDPTGALTEDSIDDTFLPVPEYINQSVPK	3	1133.544	15
pY-1069	[Y]SSDPTGALTEDSIDDTFLPVPEYINQSVPK	3	1160.199	20
pY-1092	YSSDPTGALTEDSIDDTFLPVPE[Y]INQSVPK	3	1160.199	25
pY-1110_miss	RPAGSVQNPV[Y]HNQPLNPAPSR	3	827.0718	25
Y-1110_miss	RPAGSVQNPVYHNQPLNPAPSR	3	800.4158	25
pY-1172	GSHQISLDNPD[Y]QQDFFPK	3	772.6708	15
Y-1172	GSHQISLDNPDYQQDFFPK	3	746.015	15
pY-1197	GSTAENAE[Y]LR	2	645.7717	20
Y-1197	GSTAENAEYLR	2	605.7886	20
ctrl_pep****	VAPQSSEFIGA	2	553.2798	15

*[Y] – tyrosine phosphorylation

**miss – miss cleavage

***double phosphorylation site

****control peptide for total EGFR expression

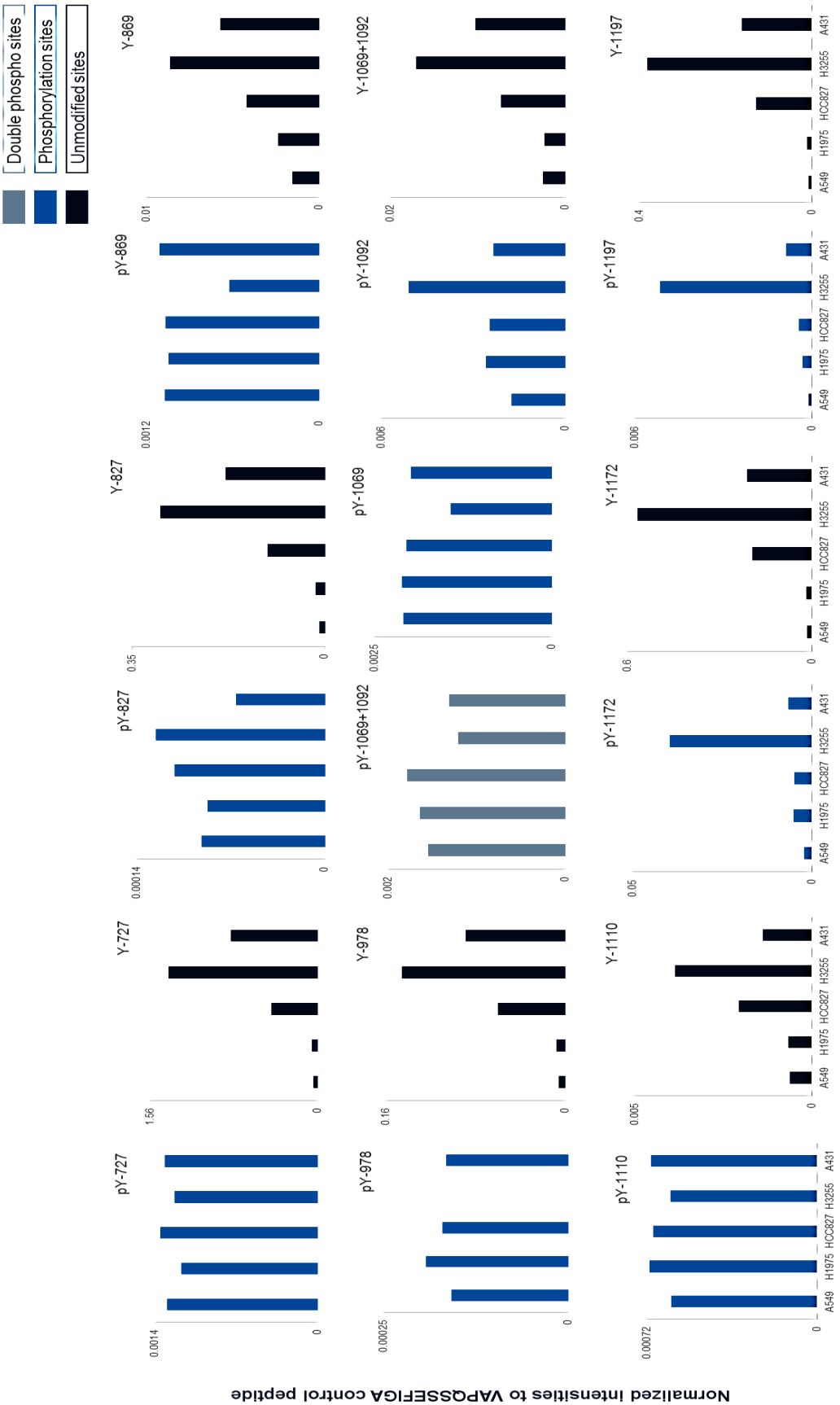


Figure IV.2. Identification of phosphorylated tyrosine sites along with the unmodified peptide sequences in the five cancer cell lines. Endogenous/heavy peptide area ratios normalized to area ratio of VAPQSSEFIGA control peptide for total EGFR expression. Expression of each phosphorylation site (pY) and its unmodified counterpart (Y) from the detected sites are presented with light and dark blue color, respectively. The double phosphorylated site at position 1069/1092 is presented with greyish-blue color (middle graph).

Phosphotyrosine single site dynamic profiling

Dynamic profiling of the phosphorylation sites in a time dependent manner upon EGF stimulation indicated when phosphotyrosine activation occurs with the highest rate. Initially, the A431 (EGFRwt) cells were starved overnight and afterwards stimulated with 100 nM EGF for 5, 7, 15 and 30 min. Different dynamic profiles were observed at each phosphorylated site. The majority of the sites were activated after 5 minutes of stimulation, whereas some sites showed slower response to EGF stimulation or loss of activity after longer stimulation. From the dynamic profiles presented in figure IV.3, we observed that the phosphotyrosine at positions 727, 869, 998, 1092 and 1110 showed the highest phosphorylation 5 minutes after stimulation, while pTyr 978, pTyr 1069 and pTyr1069/pTyr1092 showed maximal activation at 7 minutes and pTyr 1172 and pTyr 1197 at 15 minutes upon stimulation. pTyr 869 and pTyr998 exhibited reactivation after 30 minutes, while pTyr1092 was reactivated after 15 minutes. In contrast, pTyr1110 was activated early and after 5 minutes started gradually losing its activity.

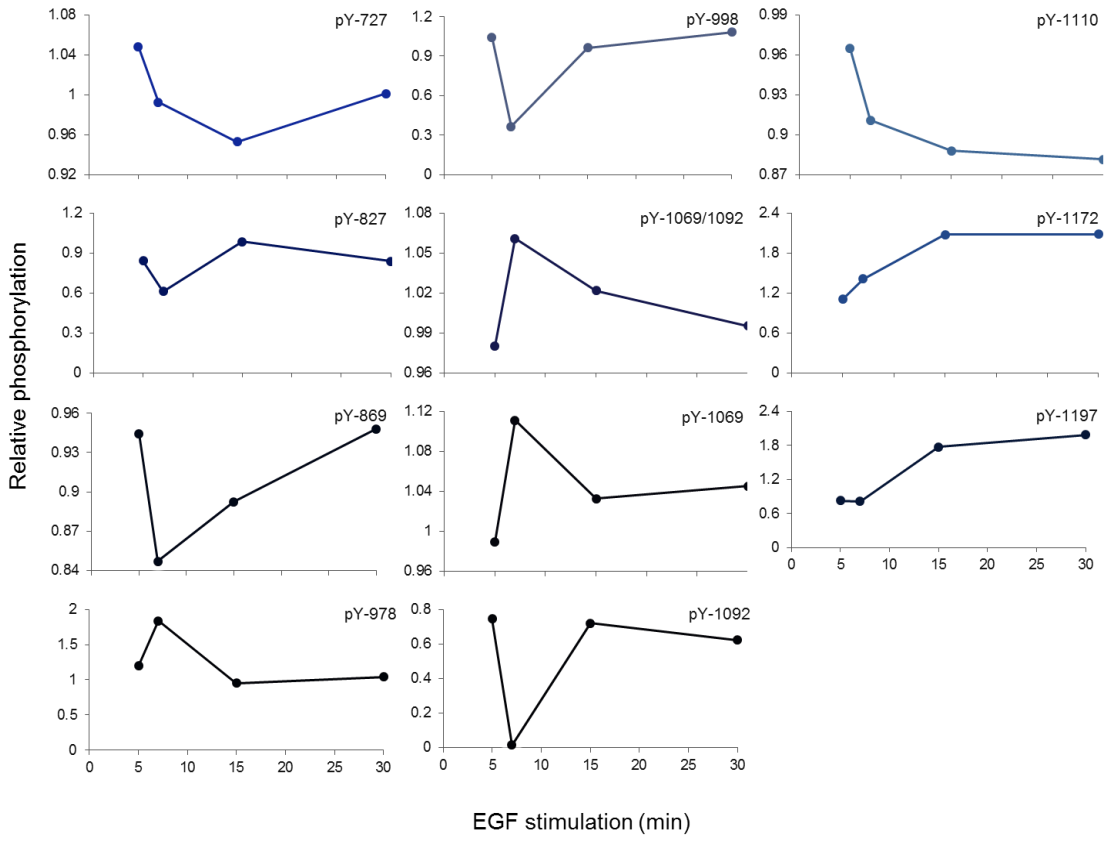


Figure IV.3. Dynamic profiles of the identified 11 phosphotyrosine peptides in A431 cells. Cells were stimulated for 5, 7, 15 and 30 min with 100 nM EGF. Each time point presented as mean from n=2 stimulated/unstimulated ratio.

Next, A549, H1975, HCC827 and A431 cells were stimulated with 200 nM EGF for 1, 3, 5, 10 and 15 since the majority of the sites were previously activated at the beginning of the stimulation. After the IP-PRM analysis, only phosphotyrosine at position 1069, 1172 and 1197 were detected among the four cell lines. Additional identification of pTyr1092 was observed in HCC827 and A431cells, cell lines overexpressing EGFR. The dynamic profiling of these sites demonstrated phosphotyrosine activation at 1 to 5 minutes in all of the cell lines (figure IV.4). Specifically, pTyr 1069 was activated after 1 minute of EGF stimulation in A549 and H1975 cells and after 3 minutes in HCC827 and A431 cell lines. The tyrosine at position 1172 showed constant activation during 15 minutes of stimulation in all cells and displayed gradual decreased activity in the H1975 cells. Finally, pTyr 1197 was activated after 3 minutes of stimulation in A549; after 3 minutes, the phosphorylation activity gradually decreased, and it was reactivated after 15 minutes. Same site showed gradual activation within the first 10 minutes, from which time point its activity dropped down. This site showed constant activation during all 15 minutes of stimulation in HCC827 and A431 cells. Also, this site presented high variability between replicates at all time points of stimulation in H1975, HCC827 and A431 cells. These results confirmed the dynamic reversible nature of the phosphorylation modification.

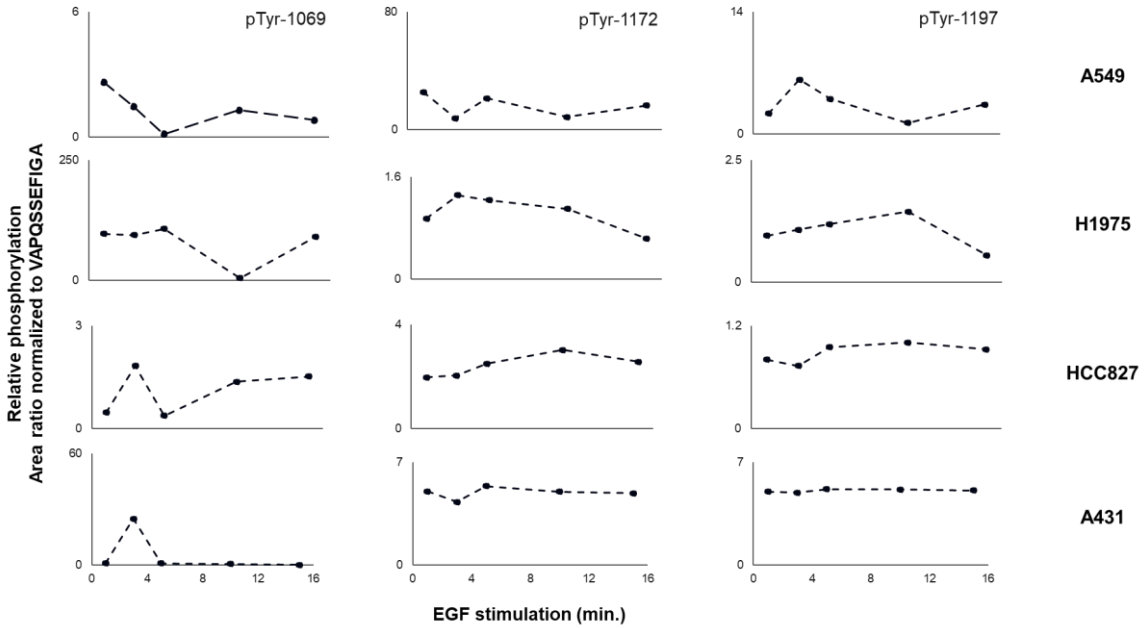


Figure IV.4. Dynamic profiles of the phosphotyrosine 1069, 1172 and 1197 peptides in A549, H1975, HCC827 and A431 cells. Cells were stimulated for 1, 3, 5, 10 and 15 min with 200 nM EGF. Each time point presented as mean from n=2 stimulated/unstimulated ratio.

Estimation of the stoichiometry of the most abundant phosphotyrosine sites

Determination of the phosphotyrosine site occupancy also known as stoichiometry of the modified fraction of the protein is an important parameter to evaluate the functionality of the phosphorylation site [197]. In these single peptide analyses, phosphotyrosine peptides, present in very low abundances, were quantified together with the unmodified counterpart using alkaline phosphate enzyme for dephosphorylation. Briefly, cells were stimulated with EGF for ten minutes before cell collection and lysis. Then, EGFR was purified from the cell lysates, digested and desalted to remove macromolecules undesirable for the MS instrument. After elution of the peptides from the desalting column, each sample was split in two; one part was directly analyzed by PRM (phosphorylated fraction), and the other fraction was dephosphorylated by alkaline phosphatase for an hour before targeted PRM analysis (dephosphorylated fraction).

Dephosphorylation yields an increase in the endogenous signal intensity of the unmodified peptide (Y) equal to the phosphorylated fraction detected with the phosphopeptide (pY) prior dephosphorylation (figure IV.5).

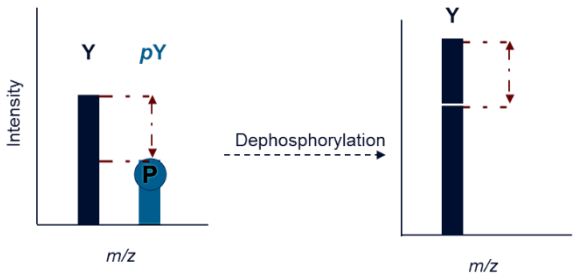


Figure IV.5. Schematic presentation of the dephosphorylation process. After alkaline phosphatase treatment, the signal intensity of the unmodified peptide increases by the phosphorylated portion before dephosphorylation.

Before targeted PRM analysis, samples were spiked with a mixture of the 10 SIL peptides; yet, only 6 phosphopeptides (pTyr1069, pTyr1092, pTyr1069/1092, pTyr1110, pTyr1172 and pTyr1197) were detected and used to calculate the site occupancy. All six phosphopeptides were identified only in the H3255 ($EGFR^{L858R}$) cell line, with the highest occupancy on pTyr1092 and pTyr1197 sites, 89% and 182%, respectively. The H1975 cell line ($EGFR^{L858R/T790M}$) had four sites occupied (pTyr1092, pTyr1110, pTyr1172 and pTyr1197) and HCC827 ($EGFR^{delE746\text{-}A750}$) and A431 (EGFRwt) showed only three site occupancies (pTyr1110, pTyr1172 and pTyr1197). A549 (EGFRwt) was the cell line with only one occupied phosphotyrosine site at position 1172. The frequencies of the detected and quantified phosphotyrosines are

shown in figure IV.5, whereas in table IV.2 the calculated dephosphorylation percentages for each site are presented.

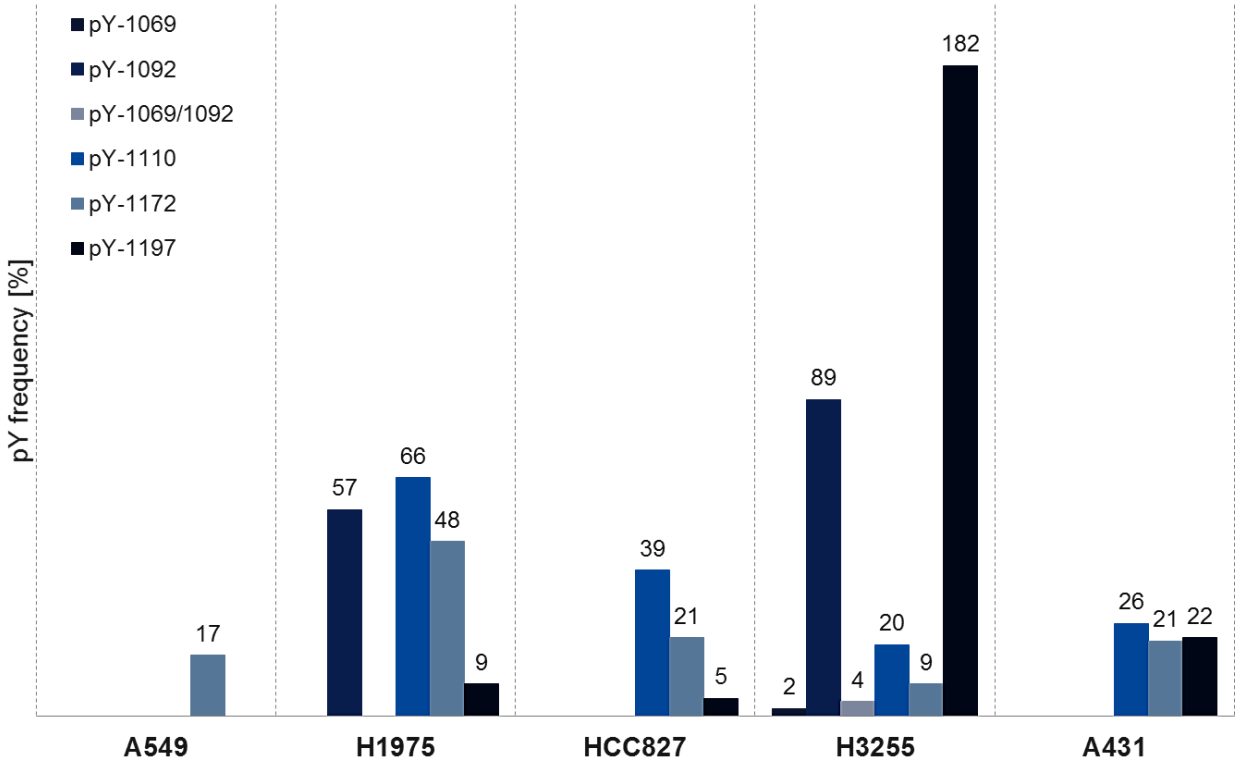


Figure IV.5. Phosphotyrosine site occupancy within each cell line estimated from endogenous/heavy peptide ratios of phosphorylated and unmodified peptides normalized to control peptide and calculated dephosphorylation rate. Six out of nine phosphopeptides were identified and quantified. Bars present as mean from n=2 replicates. Each phosphotyrosine site is presented in different blue shade.

Table IV.2. Dephosphorylation efficiency for each phosphotyrosine site.

Cell line/Tyrosine site	Tyr1069	Tyr1092	Tyr1069/1092	Tyr1110	Tyr1172	Tyr1197
A549					97%	
H1974		99%		98%	99%	100%
HCC827				96%	97%	100%
H3255	89%	99%	92%	100%	100%	100%
A431				100%	99%	100%

DISCUSSION

Protein phosphorylation analysis as defined by C E Parker *et al*. represents “a “project” rather than a routine analysis” [198]. The measurement of this highly dynamic event, which can occur on various sites in the protein sequence, requires coupling of several techniques for phosphoprotein (or phosphopeptide) isolation and sequencing with high sensitivity, specificity and dynamic range [124]. During the past several years, mass spectrometry-based analyses have been used for precise determination of the phosphorylated sites and their stoichiometry by measuring the molecular mass of phosphorylated and unmodified proteins [199, 200]. Respectively, we present here the identification of phosphotyrosine sites along the entire EGFR protein sequence and the estimation of their occupancy rate in relation to the EGFR mutational status by an IP-PRM approach in five cancer cell lines. The tyrosine phosphorylation rate in the cells is very low and thus, for detection of all phosphorylated tyrosine sites and their later characterization, EGFR was purified at protein level. Protein purification increases the sequence coverage and the identification of low abundant phosphotyrosine sites, compared to affinity peptide enrichment strategies [201]. Isolating the phosphoprotein allowed a broad phosphorylation characterization from micrograms of purified EGFR using only one pan-antibody. Our IP-PRM approach was able to detect 55% (11/20) of EGFR’s phosphotyrosine sites, while others using multiple phosphopeptide enrichment steps reported only few [202, 203].

Phosphotyrosine sites were detected and later quantified using synthetic pairs of phosphorylated and unmodified peptides with matching chemical properties, chromatographic retention time and fragmentation distribution to the endogenous peptides in the MS spectra. The main drawback of the identification process was the decreased ionization efficiency of the phosphopeptides [204] coming from the low abundancy and from low stability of the phosphoryl group, which led to the missed identification of some phosphotyrosine sites [195, 205]. On the other hand, the synthesis of labelled peptides was limited to only eleven sites due to the unsuitable length and chemical properties (amino acid hydrophobicity, coupling, cleavage, level of purification, multiple charge states etc.) of some of the peptides which decreased their detection level. However, not all the SIL peptides obtained presented a stable ionization efficiency, which resulted in the identification of only nine phosphotyrosine sites.

The applied CID fragmentation efficiently produced fragmentation profiles for the nine phosphotyrosine peptides, although electron transfer dissociation (ETD) is considered to allow higher phosphopeptide identification [195, 206]. From the tyrosine sites located in the TK domain of EGFR, the majority of which without known interaction partners, Tyr727, Tyr827 and Tyr869 were identified. The phosphotyrosine sites when activated bind to various adaptor proteins via protein-binding motifs and subsequently activate different signal transduction pathways. Tyr727 can interact with Src homology 2 domain (SH2) adaptor proteins [207]. A study comparing EGFR phosphorylation and mutation status indicated increased levels of pTyr727 in samples harboring the EGFR deletion mutation compared to the point mutation [208]. However, our study showed similar expression of this site in all the cells independently of their mutation status. Another study showed decreased phosphorylation on the Tyr869 site in A549 (EGFRwt) upon erlotinib treatment [62, 202]. This site is strongly correlated with the EGFR L858R mutation and thus considered important in cancer cells [209]. However, we detected decreased levels of pTyr869 in H3255 ($EGFR^{L858R}$) cells contradicting its relation to the point mutation.

The majority of EGFR phosphorylation studies correlate tyrosine autophosphorylation (responsible for the activation of signaling pathways) with EGFR mutational status and targeted therapies [210]. This includes the autophosphorylation of tyrosine sites from the EGFR regulatory domain. Our study identified 6 singly phosphorylated and one doubly phosphorylated tyrosine sites in this region (Tyr 978, Tyr 1069, Tyr 1092, Tyr1069/Tyr1092, Tyr1110, Tyr1172 and Tyr1197). Tyr978 is a specific docking site for interaction with STAT5 (latent cytoplasmic transcription factor) [207, 211]; however we were not able to identify this site in H3255 cells, overexpressing mutant EGFR. Tyr1069 and Tyr1092 are interacting partners with growth factor receptor-bound protein 1 (Grb1), involved in the PI3K/AKT pathway, but also exhibit a strong binding to casitas B-lineage lymphoma (Cbl) ubiquitin protein ligase [212]. The study demonstrated a close relationship between phosphorylation and ubiquitination, as the presence of Tyr1069 and Tyr1092 in mutant EGFR led to ubiquitination dose-response curves different than those of EGFRwt after 2 min stimulation. Also, both singly phosphorylated sites and their doubly phosphorylated version were activated in the first minutes upon EGF stimulation but started decreasing after 5 min; however, upon stimulation, Tyr1092, the most abundant of all three isoforms decreased in time, whereas Tyr1069 and Tyr1069/Tyr1092 strongly increased, indicating that Tyr1069 phosphorylation is causing rapid phosphorylation of Tyr1092, converting this site into a

doubly phosphorylated form [213]. Respectively, our dynamic profiling investigation using A431 (EGFRwt) cells demonstrated a similar pattern of phosphorylation behavior for Tyr1069 and Tyr1069/Tyr1092 (both increasing rapidly in the first 7 minutes and then gradually decreasing) compared to Tyr1092 (activated in the first 5 min, showing a rapid loss of activation and then reactivation after 15 minutes upon stimulation). Furthermore, Tyr1092 was considered as a predictive marker for screening patients carrying EGFRwt due to the weak correlation with EGFR mutations [189, 214]. On contrary, we found that Tyr1069 and Tyr1069/Tyr1092 expression levels were lower in the mutant H3255 cells, whereas Tyr1092 had the highest frequency compared to the other cells and thus indicated relation to the EGFR L858R mutation. The three other autophosphorylation sites, Tyr1110 (binding Grb2 involved in the MAPK/ERK pathway), Tyr1172 and Tyr1197 (both binding Shc also involved in the MAPK/ERK pathway), were found to be correlated with the EGFR sensitizing mutations and sensitivity to erlotinib [194, 215], exhibiting fast autophosphorylation events compared to wild-type EGFR. In the absence of mutations, autophosphorylation first happened at Tyr1197, then Tyr1172 and Tyr1110 [62]. Our IP-PRM analysis showed similar expression of Tyr1110 in all the cell lines, so no correlation was found with the EGFR mutation status, while Tyr1172 and Tyr1197 had the highest expression in H3255 ($EGFR^{L858R}$) cells, demonstrating a strong relationship with the point mutation. Tyr1172 and Tyr1197 were found to be inhibited upon erlotinib treatment and consequently considered as potential biomarkers for TKI sensitivity [216]. Our results support this statement, since both tyrosine sites were expressed the most in the H3255 cells, cells sensitive to erlotinib.

Establishment of phosphorylation occupancy is also important for understanding the cellular regulation mechanism, *i.e.,* the phosphorylation frequency rate is associated with the kinase/phosphatase activity and protein abundance [217]. To quantify the phosphorylation rate, the endogenous/heavy ratio of the phosphorylated peptide is compared to the endogenous/heavy ratio of the unmodified peptide considering the dephosphorylation rate in the calculations. One common limitation of the phosphoproteomics measurements is the inability to distinguish between phosphorylation sites within the same peptide as they yield chromatographic peaks with the same retention time and *m/z* [194]. However, our high resolution and accurate mass PRM analysis successfully distinguished between Tyr1069, Try1092, Tyr1069/Tyr1092 and the unmodified phosphopeptide, producing signals at four different retention times (yet the same *m/z* values for singly phosphorylated Tyr1069 and Tyr1092 peptides). The quantitative IP-PRM analysis detected the six phosphotyrosine

peptides associated with the EGFR mutation status and TKIs sensitivity. The six peptides were only all quantified in the H3255 ($EGFR^{L858R}$) cell line, where Tyr1197 and Tyr1092 were phosphorylated the most. In the H1975 cell line, also harboring the L858R mutation, four sites were phosphorylated with 9% to 66% frequency, proving the strong relationship of the autophosphorylation sites to the EGFR point mutation. In the HCC827 ($EGFR^{delE746-A750}$) cell line Try1110, Tyr1172 and Tyr1197 showed a frequency below 40%, despite the better response rate of the deletion mutation to TKIs therapies compared to patients harboring the EGFR L858R mutation [218]. The same three phosphotyrosine sites were detected in A431 (EGFRwt) cells with a frequency rate below 27% and only the Tyr1172 phosphopeptide was detected in the A549 (EGFRwt) with 17%, two cell lines harboring EGFR wild-type. Another important issue during estimation of the phosphorylation stoichiometry is the ability to differentiate the phosphorylated portion of the protein from the total protein abundance [219]. The usage of synthetically labelled peptides as internal standards helped in the unambiguous identification of each phosphotyrosine site and quantification of the stoichiometry. Respectively, this analysis demonstrated that the cells overexpressing EGFR were not simultaneously expressing the highest level of phosphorylation on specific sites.

Over the last years, although many approaches have been applied for the identification of phosphorylated sites and estimation of their stoichiometry, only a small fraction of information with clinical significance has been obtained [220]. The level of identified phosphorylation sites and their characterization is still quite low. Some key limitations of current phosphoproteomics strategies are the inability to distinguish between isoforms with multiple phosphorylation sites, the low ionization efficiency preventing correct stoichiometry estimation, and the choice of a suitable peptide enrichment method to obtain an appropriate amount of low abundant phosphopeptides [221]. In our study, we also wrestled with the unambiguous identification of low abundant phosphotyrosine sites from the complex background due to the low ionization efficiency of some peptides. One of our main challenges was achieving good reproducibility between sample replicates and experiments, proving the reversible and dynamic nature of this post-translational modification [222]. Nevertheless, our IP-PRM approach managed to successfully identify eleven key phosphotyrosine sites in a single experiment and asses the stoichiometry of the six most clinically relevant autophosphorylation sites in the EGFR protein sequence from only nano-gram amount of purified EGFR protein.

MATERIALS AND METHODS

Chemicals and Reagents

The biological material, antibody, affinity support and all other reagents used were the same as described in Chapter I. Sequencing grade modified trypsin (Promega) was used for generation of phosphotyrosine peptides and their unmodified counterparts. Phosphatase, Alkaline from bovine intestinal mucosa (Cat. No. P0114-10KU, Sigma Aldrich) was used for dephosphorylation. Epidermal growth factor (EGF) (Cat. No. 01-407, MERCK) was used for cell stimulation.

Sample preparation step

Biological material: The five cancer cell lines harboring different EGFR mutations described in Chapter I were used as biological material for the phosphorylation analyses.

Cell lysis and protein extraction: For phosphotyrosine identification, two million cells of each cell line were lysed as described in Chapter I. For the dynamic profiling and stoichiometry estimation analyses one million cells were harvested and split in serum-free media into 6-well plates and starved overnight. On the next day, EGF stimulation (100 nM or 200 nM) was carried out at different time points and cells were immediately lysed on ice and collected by scraping.

Immunopurification: EGFR protein purification was performed as described Chapter I, using the mAb supplied by Novus and protein A/G micro-columns for immunopurification on the 96-well plate automated liquid handler Versette® working station (Thermo).

Digestion, IS spiking, desalting, dephosphorylation: Purified EGFR samples were digested as described in Chapter I, except trypsin was used for peptide generation instead of GluC endopeptidase. After digestion, samples were desalted onto a 96-well plate desalting station (Waters). Regarding the identification and dynamic profiling analyses, samples were dried after desalting and re-suspended in an internal standard mix in 0.1% formic acid in water. For the quantitative analyses, samples were split in two after desalting; one fraction was vacuum dried and re-suspended in IS mix, and the second fraction was vacuum dried, then re-suspended in alkaline phosphatase buffer, and treated with 1000 u/sample of alkaline phosphatase for dephosphorylation. The dephosphorylation process was stopped

by heating the samples at 90°C for ten minutes; additional desalting, vacuum drying and re-suspension in IS mix was performed.

LC-MS targeted analysis

LC separation: The chromatographic separation of the phosphotyrosine and unmodified peptides was performed on the Dionex instrument as described in Chapter I, with only a difference in the gradient duration. Separation was done by a linear gradient starting from 2% to 90% B in 76 min.

PRM analysis: See Chapter I.

Data processing: See Chapter I.

Conclusion and Outlook

CONCLUSION AND OUTLOOK

Modified proteins as a result of the altered genes can be found in various abundances and forms in tumor cells within a broad dynamic range. These protein variations may occur due to alternative splicing, polymorphisms or post-translational modifications and can be involved in different biological processes, including cell differentiation, proliferation, signal transduction and/or apoptosis. These modifications may rise as somatic mutations (DNA alterations in the body cells) and/or as different protein isoforms (proteins with similar sequences and function). Since they may be considered as drug targets their unambiguous examination is necessary to better understand disease processes and reactions. Therefore, effective and accurate methods are required to distinguish between these altered forms that may have similar amino acid sequences and thus behave in the same way during the analysis. Moreover, their quantification is also necessary due to their presence in various concentration ranges.

A multiplexed strategy was developed for somatic “driver” mutation and protein isoform targeted analyses in low nano-gram amounts in different cancer cells. Coupling two different proteomics’ technologies, targeted MS with protein immunopurification, allowed complete analysis of targeted proteins and their isoforms. The advantages of this merger were decreased sample complexity and high-throughput analysis, where all the targets were identified in a high sensitive and selective fashion. The immunopurification allowed usage of short chromatographic separations, while the high resolution accurate PRM analysis advanced the identification of the targets, both yielding rapid analysis. Once established, this approach was used for various cancer analyses and resulted in establishing a tumor screening platform for protein based molecular diagnostic, providing results in a relatively short time (less than a week) [119].

As already published, this methodology proved its applicability to various sample types, such as plasma samples, cell lysates and tissues. Using an automated approach for protein purification, 96 different samples can be analyzed simultaneously, providing information about 96 different patients and hence assisting with patient selection [118]. On the other hand, the PRM mode measured in a single analysis all peptides carrying the mutation and the wild-type counterparts, where estimation of the expression differences between the mutation versus the wild-type was completed as an important parameter for a personalized approach.

Another benefit of this strategy is that along with the mutation profiling, characterization of the post-translational modification can be performed. The EGFR phosphorylation was explained in more details, regarding the identification of important autophosphorylation sites involved in activation of signaling pathways characteristic for the tumor initiation and progression. The phosphorylation study demonstrated the benefits of the protein purification, since the phosphotyrosine sites were better preserved and permitted identification of the majority of the activated sites compared to studies where additional peptide enrichment steps were used and only a few EGFR phosphotyrosine sites were detected [194]. Moreover, the usage of specific antibodies against different phosphorylation sites was omitted, presenting this approach as a simpler, less expensive and time-consuming method for post-translational modification characterization.

Currently the biomedical research is focused on various events in the nucleotide sequence of the DNA, different RNA processes and the post-translational modifications. The whole genome sequencing enabled classification of various DNA mutations and polymorphisms that may transcribed and translated into different protein forms. Moreover, the protein isoforms (like KRas, NRas and HRas) can also occur due to alternative splicing and RNA translation malfunctions, increasing the diversity of the proteome. At the end of the day proteins are the key elements of the cellular behavior and functions. However, genomic and proteomic fields are highly complementary and dependent on each other. Therefore, the goal of this project was to provide a scientific foundation of protein isoforms using the latest technology in mass spectrometry with translation of current knowledge on genetic mutations in tumors. The assessment of the predictive parameters in a complex tumorous specimen with limited mutant material available is a challenge. Moreover, the molecular profiling of a tumor implicates screening of individual genes for the occurrence of “driver” mutation for prediction of the targeted therapies, directed against the modified protein. Therefore, when genomic, transcriptomic and proteomic methods were compared, it pointed out the importance of performing the analysis at the protein level to obtain a clear picture of the disease state. It furthermore established the advantages of the mass spectrometry as a powerful tool to deliver information related to the proteins of interest and their modifications without any prior knowledge.

The methodologies used in the biomedical research must be analytically validated and the scientific value to be translated into a clinical environment. The validation includes testing

for precision, accuracy, dynamic ranges and comparison to existing methods. For protein analysis Western blot and ELISA are traditionally used. Both techniques use antibodies towards the targeted protein and/or peptides. Even though both methods are easy to use and give reproducible results in a short time, they have limitations. The main drawback is the antibody selectivity towards the targeted proteins, the ability to distinguish between isoforms as well as the availability of suitable antibodies and the batch-to-batch differences that decrease the reproducibility and repeatability. On contrary, MS-based approaches can provide unambiguous detection and quantification of the proteins of interest and can overcome the limitations of the antibody-based techniques, like IP, which besides for protein enrichment was used as a cleaning step prior targeted analysis.

The main advantage of MS is the ability to deliver information regarding the mutation status, isoform differentiation and present PTMs without using antibodies and prior knowledge of the target, whereas traditional methods are not sensitive enough and only provide information for the total protein concentration [75]. However, this is not always a case, especially when targeting low abundant proteins in complex tissue samples. The role of the biological matrix can significantly suppress the protein target resulting in reduced detection coverage [223] and thus requiring sample clean-up compatible with the subsequent MS analysis [224]. Nonetheless, the clinical adoption of MS is still slow due to the assay complexity and cost (instrumentation, reagents and turnaround time). With simplification of the sample preparation steps, the multiplex and high-throughput capability and the delivery of (new) clinically significant information regarding the cancer related protein isoforms can bring MS-based methods into clinical practice. This can be especially important for tumor tissue examinations with limited material for screening and testing, where all the information regarding the existing alterations at protein level can be obtained with single MS assay [225].

On this account, a screening platform may be developed for the analysis of predictive biomarkers by establishing proteomics assays. These assays applied to cancer samples might provide classification of tumor types at the molecular level instead on mainly histopathological information. Also, the information may be directly used for better choices for therapeutic intervention of patients resulting in optimal treatment.

REFERENCES

REFERENCES

1. Pikor, L., et al., *The detection and implication of genome instability in cancer.* Cancer Metastasis Reviews, 2013. **32**(3-4): p. 341-352.
2. Sun, X.-x. and Q. Yu, *Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment.* Acta Pharmacologica Sinica, 2015. **36**(10): p. 1219-1227.
3. Sever, R. and J.S. Brugge, *Signal Transduction in Cancer.* Cold Spring Harbor Perspectives in Medicine, 2015. **5**(4): p. a006098.
4. Yu, C.C., et al., *Mutant allele specific imbalance in oncogenes with copy number alterations: Occurrence, mechanisms, and potential clinical implications.* Cancer Lett, 2017. **384**: p. 86-93.
5. Cheng, J., et al., *Pan-cancer analysis of homozygous deletions in primary tumours uncovers rare tumour suppressors.* Nature Communications, 2017. **8**(1): p. 1221.
6. Morris, L.G.T. and T.A. Chan, *Therapeutic Targeting of Tumor Suppressor Genes.* Cancer, 2015. **121**(9): p. 1357-1368.
7. Romero-Laorden, N. and E. Castro, *Inherited mutations in DNA repair genes and cancer risk.* Curr Probl Cancer, 2017. **41**(4): p. 251-264.
8. Torgovnick, A. and B. Schumacher, *DNA repair mechanisms in cancer development and therapy.* Frontiers in Genetics, 2015. **6**: p. 157.
9. Stratton, M.R., P.J. Campbell, and P.A. Futreal, *The cancer genome.* Nature, 2009. **458**(7239): p. 719-24.
10. Martincorena, I. and P.J. Campbell, *Somatic mutation in cancer and normal cells.* Science, 2015. **349**(6255): p. 1483-9.
11. Torkamani, A. and N.J. Schork, *Prediction of cancer driver mutations in protein kinases.* Cancer Res, 2008. **68**(6): p. 1675-82.
12. Bozic, I., et al., *Accumulation of driver and passenger mutations during tumor progression.* Proceedings of the National Academy of Sciences, 2010. **107**(43): p. 18545-18550.
13. Pon, J.R. and M.A. Marra, *Driver and passenger mutations in cancer.* Annu Rev Pathol, 2015. **10**: p. 25-50.

14. Bignell, G.R., et al., *Signatures of mutation and selection in the cancer genome.* Nature, 2010. **463**(7283): p. 893-8.

15. Kamburov, A., et al., *Comprehensive assessment of cancer missense mutation clustering in protein structures.* Proceedings of the National Academy of Sciences, 2015. **112**(40): p. E5486-E5495.

16. Podlaha, O., et al., *Evolution of the cancer genome.* Trends Genet, 2012. **28**(4): p. 155-63.

17. Thomas, A., et al., *From targets to targeted therapies and molecular profiling in non-small cell lung carcinoma.* Annals of Oncology, 2013. **24**(3): p. 577-585.

18. Kalia, M., *Biomarkers for personalized oncology: recent advances and future challenges.* Metabolism, 2015. **64**(3 Suppl 1): p. S16-21.

19. Syn, N.L., et al., *Evolving landscape of tumor molecular profiling for personalized cancer therapy: a comprehensive review.* Expert Opin Drug Metab Toxicol, 2016. **12**(8): p. 911-22.

20. MacConaill, L.E., et al., *Clinical implementation of comprehensive strategies to characterize cancer genomes: opportunities and challenges.* Cancer Discov, 2011. **1**(4): p. 297-311.

21. Milella, M., et al., *EGFR molecular profiling in advanced NSCLC: a prospective phase II study in molecularly/clinically selected patients pretreated with chemotherapy.* J Thorac Oncol, 2012. **7**(4): p. 672-80.

22. La Thangue, N.B. and D.J. Kerr, *Predictive biomarkers: a paradigm shift towards personalized cancer medicine.* Nat Rev Clin Oncol, 2011. **8**(10): p. 587-96.

23. Duffy, M.J. and J. Crown, *A personalized approach to cancer treatment: how biomarkers can help.* Clin Chem, 2008. **54**(11): p. 1770-9.

24. Sholl, L., *Molecular diagnostics of lung cancer in the clinic.* Transl Lung Cancer Res, 2017. **6**(5): p. 560-569.

25. Shibata, T., *Current and future molecular profiling of cancer by next-generation sequencing.* Jpn J Clin Oncol, 2015. **45**(10): p. 895-9.

26. Unger, F.T., I. Witte, and K.A. David, *Prediction of individual response to anticancer therapy: historical and future perspectives.* Cellular and Molecular Life Sciences, 2015. **72**: p. 729-757.

27. Aggarwal, C., N. Somaiah, and G.R. Simon, *Biomarkers with predictive and prognostic function in non-small cell lung cancer: ready for prime time?* J Natl Compr Canc Netw, 2010. **8**(7): p. 822-32.

28. Wu, S.G., et al., *The mechanism of acquired resistance to irreversible EGFR tyrosine kinase inhibitor-afatinib in lung adenocarcinoma patients.* Oncotarget, 2016. **7**(11): p. 12404-13.

29. Koussounadis, A., et al., *Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system.* Scientific Reports, 2015. **5**: p. 10775.

30. Vogel, C. and E.M. Marcotte, *Insights into the regulation of protein abundance from proteomic and transcriptomic analyses.* Nature reviews. Genetics, 2012. **13**(4): p. 227-232.

31. Parasido, E.M., et al., *Protein drug target activation homogeneity in the face of intra-tumor heterogeneity: implications for precision medicine.* Oncotarget, 2017. **8**(30): p. 48534-48544.

32. Zugazagoitia, J., et al., *The new IASLC/ATS/ERS lung adenocarcinoma classification from a clinical perspective: current concepts and future prospects.* J Thorac Dis, 2014. **6**(Suppl 5): p. S526-36.

33. Davidson, M.R., A.F. Gazdar, and B.E. Clarke, *The pivotal role of pathology in the management of lung cancer.* Journal of Thoracic Disease, 2013. **5**(Suppl 5): p. S463-S478.

34. Chan, B.A. and B.G.M. Hughes, *Targeted therapy for non-small cell lung cancer: current standards and the promise of the future.* Translational Lung Cancer Research, 2015. **4**(1): p. 36-54.

35. Tang, J., et al., *Erlotinib Resistance in Lung Cancer: Current Progress and Future Perspectives.* Frontiers in Pharmacology, 2013. **4**: p. 15.

36. Silva, A.P.S., et al., *Targeted therapies for the treatment of non-small-cell lung cancer: Monoclonal antibodies and biological inhibitors.* Human Vaccines & Immunotherapeutics, 2017. **13**(4): p. 843-853.

37. Moumtzi, D., et al., *Prognostic factors for long term survival in patients with advanced non-small cell lung cancer.* Annals of Translational Medicine, 2016. **4**(9): p. 161.

38. Zhu YC, X.C., Ye XQ, Yin MX, Zhang JX, Du KQ, Zhang ZH, Hu J, *Lung cancer with concurrent EGFR mutation and ROS1 rearrangement: a case report and review of the literature.* OncoTargets and Therapy, 2016 **2016:9**: p. 4301—4305.

39. The Cancer Genome Atlas Research, N., *Comprehensive molecular profiling of lung adenocarcinoma.* Nature, 2014. **511**: p. 543.

40. Logue, J.S. and D.K. Morrison, *Complexity in the signaling network: insights from the use of targeted inhibitors in cancer therapy.* Genes Dev, 2012. **26**(7): p. 641-50.

41. Salgia, R., et al., *Personalized treatment of lung cancer.* Semin Oncol, 2011. **38**(2): p. 274-83.

42. Jan, et al., *Clinical Relevance of KRAS in Human Cancers.* Journal of Biomedicine and Biotechnology, 2010. **2010**: p. 13.

43. Parsons, B.L. and M.B. Myers, *KRAS mutant tumor subpopulations can subvert durable responses to personalized cancer treatments.* Personalized medicine, 2013. **10**(2): p. 191-199.

44. Castellano, E. and E. Santos, *Functional Specificity of Ras Isoforms: So Similar but So Different.* Genes & Cancer, 2011. **2**(3): p. 216-231.

45. Fernández-Medarde, A. and E. Santos, *Ras in Cancer and Developmental Diseases.* Genes & Cancer, 2011. **2**(3): p. 344-358.

46. D'Arcangelo, M. and F. Cappuzzo, *K-Ras Mutations in Non-Small-Cell Lung Cancer: Prognostic and Predictive Value.* ISRN Molecular Biology, 2012. **2012**: p. 837306.

47. Karachaliou, N., et al., *KRAS mutations in lung cancer.* Clin Lung Cancer, 2013. **14**(3): p. 205-14.

48. Linardou, H., et al., *Assessment of somatic k-RAS mutations as a mechanism associated with resistance to EGFR-targeted agents: a systematic review and meta-analysis of studies in advanced non-small-cell lung cancer and metastatic colorectal cancer.* Lancet Oncol, 2008. **9**(10): p. 962-72.

49. Irmer, D., J.O. Funk, and A. Blaukat, *EGFR kinase domain mutations - functional impact and relevance for lung cancer therapy.* Oncogene, 2007. **26**(39): p. 5693-701.

50. Siegelin, M.D. and A.C. Borczuk, *Epidermal growth factor receptor mutations in lung adenocarcinoma.* Lab Invest, 2014. **94**(2): p. 129-37.

51. Sharma, S.V., et al., *Epidermal growth factor receptor mutations in lung cancer.* Nat Rev Cancer, 2007. **7**(3): p. 169-81.

52. Gupta, R., et al., *Evaluation of EGFR abnormalities in patients with pulmonary adenocarcinoma: the need to test neoplasms with more than one method.* Mod Pathol, 2009. **22**(1): p. 128-33.

53. Gotoh, N., *Somatic mutations of the EGF receptor and their signal transducers affect the efficacy of EGF receptor-specific tyrosine kinase inhibitors.* International Journal of Clinical and Experimental Pathology, 2011. **4**(4): p. 403-409.

54. Raparia, K., et al., *Molecular profiling in non-small cell lung cancer: a step toward personalized medicine.* Arch Pathol Lab Med, 2013. **137**(4): p. 481-91.

55. Jorge, S., S.S. Kobayashi, and D.B. Costa, *Epidermal growth factor receptor (EGFR) mutations in lung cancer: preclinical and clinical data.* Brazilian Journal of Medical and Biological Research, 2014. **47**(11): p. 929-939.

56. Metro, G. and L. Crinò, *Advances on EGFR mutation for lung cancer.* Translational Lung Cancer Research, 2012. **1**(1): p. 5-13.

57. Sullivan, I. and D. Planchard, *Next-Generation EGFR Tyrosine Kinase Inhibitors for Treating EGFR-Mutant Lung Cancer beyond First Line.* Frontiers in Medicine, 2016. **3**: p. 76.

58. Antonicelli, A., et al., *EGFR-targeted therapy for non-small cell lung cancer: focus on EGFR oncogenic mutation.* Int J Med Sci, 2013. **10**(3): p. 320-30.

59. Remon, J., et al., *Acquired resistance to epidermal growth factor receptor tyrosine kinase inhibitors in EGFR-mutant non-small cell lung cancer: a new era begins.* Cancer Treat Rev, 2014. **40**(1): p. 93-101.

60. Chen, T.-C., et al., *Protein Phosphorylation Profiling Using an In Situ Proximity Ligation Assay: Phosphorylation of AURKA-Elicited EGFR-Thr654 and EGFR-Ser1046 in Lung Cancer Cells.* PLOS ONE, 2013. **8**(3): p. e55657.

61. Begley, M.J., et al., *EGF-receptor specificity for phosphotyrosine-primed substrates provides signal integration with Src.* Nature Structural &Amp; Molecular Biology, 2015. **22**: p. 983.

62. Kim, Y., et al., *Temporal resolution of autophosphorylation for normal and oncogenic forms of EGFR and differential effects of gefitinib.* Biochemistry, 2012. **51**(25): p. 5212-22.

63. Garcia, A., et al., *Performance Assessment of Epidermal Growth Factor Receptor Gene Sequencing According to Sample Size in Daily Practice Conditions.* Appl Immunohistochem Mol Morphol, 2017.

64. Strom, S.P., *Current practices and guidelines for clinical next-generation sequencing oncology testing.* Cancer Biology & Medicine, 2016. **13**(1): p. 3-11.

65. Vidal, L., et al., *Fluorescence in situ hybridization gene amplification analysis of EGFR and HER2 in patients with malignant salivary gland tumors treated with lapatinib.* Head Neck, 2009. **31**(8): p. 1006-12.

66. Hitij, N.T., et al., *Immunohistochemistry for EGFR Mutation Detection in Non-Small-Cell Lung Cancer.* Clin Lung Cancer, 2017. **18**(3): p. e187-e196.

67. Lopez-Rios, F., et al., *Comparison of molecular testing methods for the detection of EGFR mutations in formalin-fixed paraffin-embedded tissue specimens of non-small cell lung cancer.* J Clin Pathol, 2013. **66**(5): p. 381-5.

68. Drabovich, A.P., E. Martinez-Morillo, and E.P. Diamandis, *Toward an integrated pipeline for protein biomarker development.* Biochim Biophys Acta, 2015. **1854**(6): p. 677-86.

69. Mesri, M., *Advances in Proteomic Technologies and Its Contribution to the Field of Cancer.* Adv Med, 2014. **2014**: p. 238045.

70. Whiteaker, J.R. and A.G. Paulovich, *Peptide immunoaffinity enrichment coupled with mass spectrometry for peptide and protein quantification.* Clinics in laboratory medicine, 2011. **31**(3): p. 385-396.

71. Crosley, L.K., et al., *Variation in protein levels obtained from human blood cells and biofluids for platelet, peripheral blood mononuclear cell, plasma, urine and saliva proteomics.* Genes & Nutrition, 2009. **4**(2): p. 95-102.

72. Edfors, F., et al., *Gene-specific correlation of RNA and protein levels in human cells and tissues.* Molecular Systems Biology, 2016. **12**(10).

73. García-Muñoz, A., et al., *The orosomucoid 1 protein (α1 acid glycoprotein) is overexpressed in odontogenic myxoma.* Proteome Science, 2012. **10**: p. 49-49.

74. Paul, D., et al., *Mass Spectrometry-Based Proteomics in Molecular Diagnostics: Discovery of Cancer Biomarkers Using Tissue Culture.* BioMed Research International, 2013. **2013**: p. 16.

75. Nedelkov, D., *Human proteoforms as new targets for clinical mass spectrometry protein tests.* Expert Rev Proteomics, 2017. **14**(8): p. 691-699.

76. Weiß, F., et al., *Catch and measure-mass spectrometry-based immunoassays in biomarker research.* Biochimica et biophysica acta, 2014. **1844**(5): p. 927-932.

77. Bustos, D., et al., *Characterizing ubiquitination sites by peptide-based immunoaffinity enrichment.* Mol Cell Proteomics, 2012. **11**(12): p. 1529-40.

78. Stokes, M.P., et al., *PTMScan direct: identification and quantification of peptides from critical signaling proteins by immunoaffinity enrichment coupled with LC-MS/MS.* Mol Cell Proteomics, 2012. **11**(5): p. 187-201.

79. Zhao, L., et al., *Quantification of proteins using peptide immunoaffinity enrichment coupled with mass spectrometry.* J Vis Exp, 2011(53).

80. Kuhn, E., et al., *Interlaboratory Evaluation of Automated, Multiplexed Peptide Immunoaffinity Enrichment Coupled to Multiple Reaction Monitoring Mass Spectrometry for Quantifying Proteins in Plasma.* Molecular & Cellular Proteomics : MCP, 2012. **11**(6): p. M111.013854.

81. Ruppen-Canas, I., et al., *An improved quantitative mass spectrometry analysis of tumor specific mutant proteins at high sensitivity.* Proteomics, 2012. **12**(9): p. 1319-27.

82. Krastins, B., et al., *Rapid development of sensitive, high-throughput, quantitative and highly selective mass spectrometric targeted immunoassays for clinically important proteins in human plasma and serum.* Clin Biochem, 2013. **46**(6): p. 399-410.

83. Yassine, H., et al., *Mass Spectrometric Immunoassay and Multiple Reaction Monitoring as Targeted MS-based Quantitative Approaches in Biomarker Development: Potential Applications to Cardiovascular Disease and Diabetes.* Proteomics. Clinical applications, 2013. **7**(0): p. 528-540.

84. MEYFOUR, A.R.T., Mostafa; SADEGHI, Mohammad Reza, *Common Proteomic Technologies, Applications, and their Limitations.* Journal of Paramedical Sciences, 2013. **v. 4**.

85. Francesco Di Girolamo, I.L., Maurizio Muraca and Lorenza Putignani, *The Role of Mass Spectrometry in the "Omics" Era.* Curr Org Chem, 2013. **17**(23): p. 2891–2905.

86. Trenchevska, O., R.W. Nelson, and D. Nedelkov, *Mass Spectrometric Immunoassays in Characterization of Clinically Significant Proteoforms.* Proteomes, 2016. **4**(1): p. 13.

87. Maiolica, A., et al., *Targeted proteome investigation via selected reaction monitoring mass spectrometry.* Journal of proteomics, 2012. **75**(12): p. 3495-3513.

88. Bourmaud, A., S. Gallien, and B. Domon, *Parallel reaction monitoring using quadrupole-Orbitrap mass spectrometer: Principle and applications.* Proteomics, 2016. **16**(15-16): p. 2146-59.

89. Rauniyar, N., *Parallel Reaction Monitoring: A Targeted Experiment Performed Using High Resolution and High Mass Accuracy Mass Spectrometry.* International Journal of Molecular Sciences, 2015. **16**(12): p. 28566-28581.

90. Peterson, A.C., et al., *Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics.* Mol Cell Proteomics, 2012. **11**(11): p. 1475-88.

91. Parker, C.E. and C.H. Borchers, *Mass spectrometry based biomarker discovery, verification, and validation--quality assurance and control of protein biomarker assays.* Mol Oncol, 2014. **8**(4): p. 840-58.

92. Hüttenhain, R., et al., *Perspectives of targeted mass spectrometry for protein biomarker verification.* Current opinion in chemical biology, 2009. **13**(5-6): p. 518-525.

93. Leitner, A., et al., *Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines.* Trends Biochem Sci, 2016. **41**(1): p. 20-32.

94. Rosati, S., et al., *In-depth qualitative and quantitative analysis of composite glycosylation profiles and other micro-heterogeneity on intact monoclonal antibodies by high-resolution native mass spectrometry using a modified Orbitrap.* mAbs, 2013. **5**(6): p. 917-924.

95. Lesur, A., et al., *Screening protein isoforms predictive for cancer using immunoaffinity capture and fast LC-MS in PRM mode.* Proteomics Clin Appl, 2015. **9**(7-8): p. 695-705.

96. Whiteaker, J.R., et al., *Peptide Immunoaffinity Enrichment with Targeted Mass Spectrometry: Application to Quantification of ATM Kinase Phospho-signaling.* Methods in molecular biology (Clifton, N.J.), 2017. **1599**: p. 197-213.

97. Miteva, Y.V., H.G. Budayeva, and I.M. Cristea, *Proteomics-based methods for discovery, quantification, and validation of protein-protein interactions.* Anal Chem, 2013. **85**(2): p. 749-68.

98. Arnold, T. and D. Linke, *Phase separation in the isolation and purification of membrane proteins.* Biotechniques, 2007. **43**(4): p. 427-30, 432, 434 passim.

99. Moser, S.M.a.A., *Affinity Chromatography: Principles and Applications, Affinity Chromatography.* Available from: http's://WWW/books/affinity-chromatography/affinity-chromatography-principles-and-applications, 2012.

100. Nelson, R.W. and C.R. Borges, *Mass spectrometric immunoassay revisited.* J Am Soc Mass Spectrom, 2011. **22**(6): p. 960-8.

101. Gundry, R.L., et al., *Preparation of Proteins and Peptides for Mass Spectrometry Analysis in a Bottom-Up Proteomics Workflow.* Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.], 2009. **CHAPTER**: p. Unit10.25-Unit10.25.

102. Prasad, B. and J.D. Unadkat, *Optimized Approaches for Quantification of Drug Transporters in Tissues and Cells by MRM Proteomics.* The AAPS Journal, 2014. **16**(4): p. 634-648.

103. Trevisiol, S., et al., *The use of proteases complementary to trypsin to probe isoforms and modifications.* Proteomics, 2016. **16**(5): p. 715-28.

104. Lesur, A. and B. Domon, *Advances in high-resolution accurate mass spectrometry application to targeted proteomics.* Proteomics, 2015. **15**(5-6): p. 880-90.

105. Nedelkov, D., *Mass spectrometry protein tests: ready for prime time (or not).* Expert Review of Proteomics, 2017. **14**(1): p. 1-7.

106. Tipton, J.D., et al., *Analysis of Intact Protein Isoforms by Mass Spectrometry.* The Journal of Biological Chemistry, 2011. **286**(29): p. 25451-25458.

107. Xiong, Y., et al., *Immunohistochemical detection of mutations in the epidermal growth factor receptor gene in lung adenocarcinomas using mutation-specific antibodies.* Diagnostic Pathology, 2013. **8**(1): p. 27.

108. Pirker, R., et al., *EGFR expression as a predictor of survival for first-line chemotherapy plus cetuximab in patients with advanced non-small-cell lung cancer: analysis of data from the phase 3 FLEX study.* Lancet Oncol, 2012. **13**(1): p. 33-42.

109. Ummanni, R., et al., *Evaluation of reverse phase protein array (RPPA)-based pathway-activation profiling in 84 non-small cell lung cancer (NSCLC) cell lines as platform for cancer proteomics and biomarker discovery.* Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 2014. **1844**(5): p. 950-959.

110. Lievre, A., et al., *Protein biomarkers predictive for response to anti-EGFR treatment in RAS wild-type metastatic colorectal carcinoma.* Br J Cancer, 2017.

111. Hale, J.E., *Advantageous Uses of Mass Spectrometry for the Quantification of Proteins.* International Journal of Proteomics, 2013. **2013**: p. 6.

112. Trenchevska, O., E. Kamcheva, and D. Nedelkov, *Mass Spectrometric Immunoassay for Quantitative Determination of Protein Biomarker Isoforms.* Journal of proteome research, 2010. **9**(11): p. 5969-5973.

113. Ayoglu, B., et al., *Multiplexed protein profiling by sequential affinity capture.* Proteomics, 2016. **16**(8): p. 1251-6.

114. Conlon, F.L., et al., *Immunoisolation of protein complexes from Xenopus.* Methods Mol Biol, 2012. **917**: p. 369-90.

115. Gallien, S. and B. Domon, *Advances in high-resolution quantitative proteomics: implications for clinical applications.* Expert Rev Proteomics, 2015. **12**(5): p. 489-98.

116. Such-Sanmartin, G., et al., *Detection and differentiation of 22 kDa and 20 kDa Growth Hormone proteoforms in human plasma by LC-MS/MS.* Biochim Biophys Acta, 2015. **1854**(4): p. 284-90.

117. Prakash, A., et al., *Interlaboratory Reproducibility of Selective Reaction Monitoring Assays Using Multiple Upfront Analyte Enrichment Strategies.* Journal of Proteome Research, 2012. **11**(8): p. 3986-3995.

118. Oran, P.E., et al., *Parallel Workflow for High-Throughput (>1,000 Samples/Day) Quantitative Analysis of Human Insulin-Like Growth Factor 1 Using Mass Spectrometric Immunoassay.* PLoS ONE, 2014. **9**(3): p. e92801.

119. Peterman, S., et al., *An automated, high-throughput method for targeted quantification of intact insulin and its therapeutic analogs in human serum or plasma coupling mass spectrometric immunoassay with high resolution and accurate mass detection (MSIA-HR/AM).* Proteomics, 2014. **14**(12): p. 1445-56.

120. Trenchevska, O., et al., *Quantitative mass spectrometric immunoassay for the chemokine RANTES and its variants.* J Proteomics, 2015. **116**: p. 15-23.

121. Ruppen-Cañás, I., et al., *An improved quantitative mass spectrometry analysis of tumor specific mutant proteins at high sensitivity.* PROTEOMICS, 2012. **12**(9): p. 1319-1327.

122. Soh, J., et al., *Oncogene mutations, copy number gains and mutant allele specific imbalance (MASI) frequently occur together in tumor cells.* PLoS One, 2009. **4**(10): p. e7464.

123. Domon, B. and R. Aebersold, *Options and considerations when selecting a quantitative proteomics strategy.* Nat Biotech, 2010. **28**(7): p. 710-721.

124. Zhang, Y., et al., *Protein Analysis by Shotgun/Bottom-up Proteomics.* Chemical reviews, 2013. **113**(4): p. 2343-2394.

125. Boja, E.S. and H. Rodriguez, *Mass spectrometry-based targeted quantitative proteomics: achieving sensitive and reproducible detection of proteins.* Proteomics, 2012. **12**(8): p. 1093-110.

126. Maus, M.K., et al., *KRAS mutations in non-small-cell lung cancer and colorectal cancer: implications for EGFR-targeted therapies.* Lung Cancer, 2014. **83**(2): p. 163-7.

127. Lee, T., et al., *Non-small Cell Lung Cancer with Concomitant EGFR, KRAS, and ALK Mutation: Clinicopathologic Features of 12 Cases.* Journal of Pathology and Translational Medicine, 2016. **50**(3): p. 197-203.

128. Benesova, L., et al., *Multiplicity of EGFR and KRAS mutations in non-small cell lung cancer (NSCLC) patients treated with tyrosine kinase inhibitors.* Anticancer Res, 2010. **30**(5): p. 1667-71.

129. Arun M Unni, W.W.L., Kreshnik Zejnullahu, Shih-Queen Lee-Lin, Harold Varmus *Evidence that synthetic lethality underlies the mutual exclusivity of oncogenic KRAS and EGFR mutations in lung adenocarcinoma.* eLIFE, 2015. **4**.

130. Newlaczyl, A.U., J.M. Coulson, and I.A. Prior, *Quantification of spatiotemporal patterns of Ras isoform expression during development.* Scientific Reports, 2017. **7**: p. 41297.

131. Hobbs, G.A., C.J. Der, and K.L. Rossman, *RAS isoforms and mutations in cancer at a glance.* Journal of Cell Science, 2016. **129**(7): p. 1287-1292.

132. Mageean, C.J., et al., *Absolute Quantification of Endogenous Ras Isoform Abundance.* PLoS ONE, 2015. **10**(11): p. e0142674.

133. Gandhi, J., et al., *Alterations in Genes of the EGFR Signaling Pathway and Their Relationship to EGFR Tyrosine Kinase Inhibitor Sensitivity in Lung Cancer Cell Lines*. Vol. 4. 2009. e4576.

134. Malapelle, U., et al., *EGFR mutant allelic-specific imbalance assessment in routine samples of non-small cell lung cancer.* J Clin Pathol, 2015. **68**(9): p. 739-41.

135. Wang, J., et al., *Quantifying EGFR Alterations in the Lung Cancer Genome with Nanofluidic Digital PCR Arrays.* Clinical Chemistry, 2010. **56**(4): p. 623-632.

136. Liu, T.-C., et al., *Role of epidermal growth factor receptor in lung cancer and targeted therapies.* American Journal of Cancer Research, 2017. **7**(2): p. 187-202.

137. Labbe, C., et al., *Prognostic and predictive effects of TP53 co-mutation in patients with EGFR-mutated non-small cell lung cancer (NSCLC).* Lung Cancer, 2017. **111**: p. 23-29.

138. Tan, S.J., et al., *Personalized Treatment Through Detection and Monitoring of Genetic Aberrations in Single Circulating Tumor Cells.* Adv Exp Med Biol, 2017. **994**: p. 255-273.

139. Shi, X., et al., *Screening for major driver oncogene alterations in adenosquamous lung carcinoma using PCR coupled with next-generation and Sanger sequencing methods.* Sci Rep, 2016. **6**: p. 22297.

140. Wang, Q., et al., *Analysis of the status of EGFR, ROS1 and MET genes in non-small cell lung adenocarcinoma.* J buon, 2017. **22**(4): p. 1053-1060.

141. Gao, J., et al., *Comparison of Next-Generation Sequencing, Quantitative PCR, and Sanger Sequencing for Mutation Profiling of EGFR, KRAS, PIK3CA and BRAF in Clinical Lung Tumors.* Clin Lab, 2016. **62**(4): p. 689-96.

142. Reynolds, J.P., et al., *Next-generation sequencing of liquid-based cytology non-small cell lung cancer samples.* Cancer, 2017. **125**(3): p. 178-187.

143. Ivanov, M., et al., *Towards standardization of next-generation sequencing of FFPE samples for clinical oncology: intrinsic obstacles and possible solutions.* J Transl Med, 2017. **15**(1): p. 22.

144. Shao, D., et al., *A targeted next-generation sequencing method for identifying clinically relevant mutation profiles in lung adenocarcinoma.* Sci Rep, 2016. **6**: p. 22338.

145. Bennett, C.W., et al., *Cell-free DNA and next-generation sequencing in the service of personalized medicine for lung cancer.* Oncotarget, 2016. **7**(43): p. 71013-71035.

146. Ellison, G., et al., *EGFR mutation testing in lung cancer: a review of available methods and their use for analysis of tumour tissue and cytology samples.* J Clin Pathol, 2013. **66**(2): p. 79-89.

147. Tatematsu, T., et al., *The detectability of the pretreatment EGFR T790M mutations in lung adenocarcinoma using CAST-PCR and digital PCR.* J Thorac Dis, 2017. **9**(8): p. 2397-2403.

148. Jain, D., et al., *Use of Exfoliative Specimens and Fine-Needle Aspiration Smears for Mutation Testing in Lung Adenocarcinoma.* Acta Cytol, 2017.

149. Bubendorf, L., et al., *Prevalence and clinical association of MET gene overexpression and amplification in patients with NSCLC: Results from the European Thoracic Oncology Platform (ETOP) Lungscape project.* Lung Cancer, 2017. **111**: p. 143-149.

150. Thunnissen, E., et al., *Immunohistochemistry of Pulmonary Biomarkers: A Perspective From Members of the Pulmonary Pathology Society.* Arch Pathol Lab Med, 2017.

151. Ping, W., et al., *Immunohistochemistry with a novel mutation-specific monoclonal antibody as a screening tool for the EGFR L858R mutational status in primary lung adenocarcinoma.* Tumour Biol, 2015. **36**(2): p. 693-700.

152. Seo, A.N., et al., *Novel EGFR mutation-specific antibodies for lung adenocarcinoma: highly specific but not sensitive detection of an E746_A750 deletion in exon 19 and an L858R mutation in exon 21 by immunohistochemistry.* Lung Cancer, 2014. **83**(3): p. 316-23.

153. Chan, K.I., et al., *Relationship between driver gene mutations, their relative protein expressions and survival in non-small cell lung carcinoma in Macao.* Clin Respir J, 2017.

154. Que, D., et al., *EGFR mutation status in plasma and tumor tissues in non-small cell lung cancer serves as a predictor of response to EGFR-TKI treatment.* Cancer Biol Ther, 2016. **17**(3): p. 320-7.

155. Li, X., et al., *Comprehensive Analysis of EGFR-Mutant Abundance and Its Effect on Efficacy of EGFR TKIs in Advanced NSCLC with EGFR Mutations.* J Thorac Oncol, 2017. **12**(9): p. 1388-1397.

156. Young, E.C., et al., *A comparison of methods for EGFR mutation testing in non-small cell lung cancer.* Diagn Mol Pathol, 2013. **22**(4): p. 190-5.

157. Seki, Y., et al., *Picoliter-Droplet Digital Polymerase Chain Reaction-Based Analysis of Cell-Free Plasma DNA to Assess EGFR Mutations in Lung Adenocarcinoma That Confer Resistance to Tyrosine-Kinase Inhibitors.* Oncologist, 2016. **21**(2): p. 156-64.

158. Heller, G., C.C. Zielinski, and S. Zochbauer-Muller, *Lung cancer: from single-gene methylation to methylome profiling.* Cancer Metastasis Rev, 2010. **29**(1): p. 95-107.

159. Heller, G., et al., *Genome-wide CpG island methylation analyses in non-small cell lung cancer patients.* Carcinogenesis, 2013. **34**(3): p. 513-21.

160. Heller, G., et al., *Genome-wide miRNA expression profiling identifies miR-9-3 and miR-193a as targets for DNA methylation in non-small cell lung cancers.* Clin Cancer Res, 2012. **18**(6): p. 1619-29.

161. Bjaanaes, M.M., et al., *Genome-wide DNA methylation analyses in lung adenocarcinomas: Association with EGFR, KRAS and TP53 mutation status, gene expression and prognosis.* Mol Oncol, 2016. **10**(2): p. 330-43.

162. Cyll, K., et al., *Tumour heterogeneity poses a significant challenge to cancer biomarker research.* Br J Cancer, 2017. **117**(3): p. 367-375.

163. Lund, H., et al., *Exploring the complementary selectivity of immunocapture and MS detection for the differentiation between hCG isoforms in clinically relevant samples.* J Proteome Res, 2009. **8**(11): p. 5241-52.

164. Abbatiello, S.E., et al., *Large-Scale Interlaboratory Study to Develop, Analytically Validate and Apply Highly Multiplexed, Quantitative Peptide Assays to Measure Cancer-Relevant Proteins in Plasma.* Mol Cell Proteomics, 2015. **14**(9): p. 2357-74.

165. Kim, Y.J., et al., *Quantification of SAA1 and SAA2 in lung cancer plasma using the isotype-specific PRM assays.* Proteomics, 2015. **15**(18): p. 3116-25.

166. Alfaro, J.A., et al., *Onco-proteogenomics: cancer proteomics joins forces with genomics.* Nat Methods, 2014. **11**(11): p. 1107-13.

167. Pao, W., et al., *EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib.* Proc Natl Acad Sci U S A, 2004. **101**(36): p. 13306-11.

168. Furuyama, K., et al., *Sensitivity and kinase activity of epidermal growth factor receptor (EGFR) exon 19 and others to EGFR-tyrosine kinase inhibitors.* Cancer Sci, 2013. **104**(5): p. 584-9.

169. Mitsudomi, T. and Y. Yatabe, *Epidermal growth factor receptor in relation to tumor development: EGFR gene and cancer.* Febs j, 2010. **277**(2): p. 301-8.

170. Hsu, K.H., et al., *Higher frequency but random distribution of EGFR mutation subtypes in familial lung cancer patients.* Oncotarget, 2016. **7**(33): p. 53299-53308.

171. Yu, Y. and J. He, *Molecular classification of non-small-cell lung cancer: diagnosis, individualized treatment, and prognosis.* Front Med, 2013. **7**(2): p. 157-71.

172. Hammerschmidt, S. and H. Wirtz, *Lung cancer: current diagnosis and treatment.* Dtsch Arztebl Int, 2009. **106**(49): p. 809-18; quiz 819-20.

173. Mao, C., et al., *Concordant analysis of KRAS, BRAF, PIK3CA mutations, and PTEN expression between primary colorectal cancer and matched metastases.* Sci Rep, 2015. **5**: p. 8065.

174. Pao, W. and N. Girard, *New driver mutations in non-small-cell lung cancer.* Lancet Oncol, 2011. **12**(2): p. 175-80.

175. Rose-James, A. and S. TT, *Molecular Markers with Predictive and Prognostic Relevance in Lung Cancer.* Lung Cancer International, 2012. **2012**: p. 12.

176. Yatabe, Y., T. Takahashi, and T. Mitsudomi, *Epidermal growth factor receptor gene amplification is acquired in association with tumor progression of EGFR-mutated lung cancer.* Cancer Res, 2008. **68**(7): p. 2106-11.

177. Liang, Z., et al., *Relationship between EGFR expression, copy number and mutation in lung adenocarcinomas.* BMC Cancer, 2010. **10**: p. 376.

178. Suda, K., et al., *Heterogeneity of EGFR Aberrations and Correlation with Histological Structures: Analyses of Therapy-Naive Isogenic Lung Cancer Lesions with EGFR Mutation.* J Thorac Oncol, 2016. **11**(10): p. 1711-7.

179. Tabara, K., et al., *Loss of activating EGFR mutant gene contributes to acquired resistance to EGFR tyrosine kinase inhibitors in lung cancer cells.* PLoS One, 2012. **7**(7): p. e41017.

180. Zarrei, M., et al., *A copy number variation map of the human genome.* Nat Rev Genet, 2015. **16**(3): p. 172-183.

181. Campomenosi, P., et al., *A comparison between quantitative PCR and droplet digital PCR technologies for circulating microRNA quantification in human lung cancer.* BMC Biotechnol, 2016. **16**(1): p. 60.

182. Taylor, S.C., G. Laperriere, and H. Germain, *Droplet Digital PCR versus qPCR for gene expression analysis with low abundant targets: from variable nonsense to publication quality data.* Scientific Reports, 2017. **7**(1): p. 2409.

183. Sheng, M., et al., *Comparison of clinical outcomes of patients with non-small-cell lung cancer harbouring epidermal growth factor receptor exon 19 or exon 21 mutations after tyrosine kinase inhibitors treatment: a meta-analysis.* Eur J Clin Pharmacol, 2016. **72**(1): p. 1-11.

184. Aebersold, R. and M. Mann, *Mass-spectrometric exploration of proteome structure and function.* Nature, 2016. **537**(7620): p. 347-55.

185. Clarke, W., J.M. Rhea, and R. Molinaro, *Challenges in implementing clinical liquid chromatography-tandem mass spectrometry methods--the light at the end of the tunnel.* J Mass Spectrom, 2013. **48**(7): p. 755-67.

186. Liebler, D.C. and L.J. Zimmerman, *Targeted Quantitation of Proteins by Mass Spectrometry.* Biochemistry, 2013. **52**(22): p. 3797-3806.

187. Hung, M.-S., et al., *The content of mutant EGFR DNA correlates with response to EGFR-TKIs in lung adenocarcinoma patients with common EGFR mutations.* Medicine, 2016. **95**(26): p. e3991.

188. Vogel, C., et al., *Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line.* Molecular Systems Biology, 2010. **6**: p. 400-400.

189. Wang, F., et al., *Phosphorylated EGFR expression may predict outcome of EGFR-TKIs therapy for the advanced NSCLC patients with wild-type EGFR.* Journal of Experimental & Clinical Cancer Research, 2012. **31**(1): p. 65.

190. Chumbalkar, V., et al., *Analysis of phosphotyrosine signaling in glioblastoma identifies STAT5 as a novel downstream target of DeltaEGFR.* J Proteome Res, 2011. **10**(3): p. 1343-52.

191. Yang, J.L., et al., *Significance of Phosphorylated Epidermal Growth Factor Receptor and Its Signal Transducers in Human Soft Tissue Sarcoma.* Int J Mol Sci, 2017. **18**(6).

192. Kanshin, E., S. Michnick, and P. Thibault, *Sample preparation and analytical strategies for large-scale phosphoproteomics experiments.* Semin Cell Dev Biol, 2012. **23**(8): p. 843-53.

193. López, E. and W.C.S. Cho, *Phosphoproteomics and Lung Cancer Research.* International Journal of Molecular Sciences, 2012. **13**(10): p. 12287-12314.

194. Zhang, G., et al., *Mass spectrometry mapping of epidermal growth factor receptor phosphorylation related to oncogenic mutations and tyrosine kinase inhibitor sensitivity.* J Proteome Res, 2011. **10**(1): p. 305-19.

195. Johnson, H. and F.M. White, *Towards Quantitative Phosphotyrosine Profiling In Vivo.* Seminars in cell & developmental biology, 2012. **23**(8): p. 854-862.

196. Nishi, H., A. Shaytan, and A.R. Panchenko, *Physicochemical mechanisms of protein regulation by phosphorylation.* Frontiers in Genetics, 2014. **5**: p. 270.

197. Olsen, J.V. and M. Mann, *Status of Large-scale Analysis of Post-translational Modifications by Mass Spectrometry.* Molecular & Cellular Proteomics : MCP, 2013. **12**(12): p. 3444-3452.

198. Parker CE, M.V., Mocanu M, et al., *Mass Spectrometry for Post-Translational Modifications.* . Neuroproteomics, 2010.

199. Sharma, K., et al., *Ultradeep Human Phosphoproteome Reveals a Distinct Regulatory Nature of Tyr and Ser/Thr-Based Signaling.* Cell Reports, 2014. **8**(5): p. 1583-1594.

200. Jin, L.L., et al., *Measurement of Protein Phosphorylation Stoichiometry by Selected Reaction Monitoring Mass Spectrometry.* Journal of Proteome Research, 2010. **9**(5): p. 2752-2761.

201. Engholm-Keller, K. and M.R. Larsen, *Technologies and challenges in large-scale phosphoproteomics.* Proteomics, 2013. **13**(6): p. 910-31.

202. Assiddiq, B.F., et al., *EGFR S1166 phosphorylation induced by a combination of EGF and gefitinib has a potentially negative impact on lung cancer cell growth.* J Proteome Res, 2012. **11**(8): p. 4110-9.

203. Dong, M., et al., *Sensitive, Robust, and Cost-Effective Approach for Tyrosine Phosphoproteome Analysis.* Anal Chem, 2017. **89**(17): p. 9307-9314.

204. Pan, N., et al., *Highly efficient ionization of phosphopeptides at low pH by desorption electrospray ionization mass spectrometry.* Analyst, 2013. **138**(5): p. 1321-1324.

205. Betts, M.J., et al., *Systematic identification of phosphorylation-mediated protein interaction switches.* PLOS Computational Biology, 2017. **13**(3): p. e1005462.

206. Collins, M.O., et al., *Confident and sensitive phosphoproteomics using combinations of collision induced dissociation and electron transfer dissociation.* Journal of Proteomics, 2014. **103**(Supplement C): p. 1-14.

207. Schulze, W.X., L. Deng, and M. Mann, *Phosphotyrosine interactome of the ErbB-receptor kinase family.* Mol Syst Biol, 2005. **1**: p. 2005.0008.

208. Guha, U., et al., *Comparisons of tyrosine phosphorylated proteins in cells expressing lung cancer-specific alleles of EGFR and KRAS.* Proc Natl Acad Sci U S A, 2008. **105**(37): p. 14112-7.

209. Kim, Y., et al., *Differential Effects of Tyrosine Kinase Inhibitors on Normal and Oncogenic EGFR Signaling and Downstream Effectors.* Mol Cancer Res, 2015. **13**(4): p. 765-74.

210. Huang, P.H., *Phosphoproteomic studies of receptor tyrosine kinases: future perspectives.* Mol Biosyst, 2012. **8**(4): p. 1100-7.

211. Shpakov, A.O., *Signal Protein-Derived Peptides as Functional Probes and Regulators of Intracellular Signaling.* Journal of Amino Acids, 2011. **2011**: p. 25.

212. Capuani, F., et al., *Quantitative analysis reveals how EGFR activation and downregulation are coupled in normal but not in cancer cells.* Nature Communications, 2015. **6**: p. 7999.

213. Curran, T.G., et al., *MARQUIS: a multiplex method for absolute quantification of peptides and posttranslational modifications.* Nat Commun, 2015. **6**: p. 5924.

214. Sette, G., et al., *Tyr1068-phosphorylated epidermal growth factor receptor (EGFR) predicts cancer stem cell targeting by erlotinib in preclinical models of wild-type EGFR lung cancer.* Cell Death Dis, 2015. **6**: p. e1850.

215. Zhang, X., et al., *Identifying novel targets of oncogenic EGF receptor signaling in lung cancer through global phosphoproteomics.* Proteomics, 2015. **15**(2-3): p. 340-55.

216. Zhang, X., et al., *Quantitative Tyrosine Phosphoproteomics of Epidermal Growth Factor Receptor (EGFR) Tyrosine Kinase Inhibitor-treated Lung Adenocarcinoma Cells Reveals Potential Novel Biomarkers of Therapeutic Response.* Mol Cell Proteomics, 2017. **16**(5): p. 891-910.

217. Wu, R., et al., *Correct interpretation of comprehensive phosphorylation dynamics requires normalization by protein expression changes.* Mol Cell Proteomics, 2011. **10**(8): p. M111.009654.

218. Yu, J.-Y., et al., *Clinical outcomes of EGFR-TKI treatment and genetic heterogeneity in lung adenocarcinoma patients with EGFR mutations on exons 19 and 21.* Chinese Journal of Cancer, 2016. **35**: p. 30.

219. Tsai, C.-F., et al., *Large-scale determination of absolute phosphorylation stoichiometries in human cells by motif-targeting quantitative proteomics.* 2015. **6**: p. 6622.

220. Junger, M.A. and R. Aebersold, *Mass spectrometry-driven phosphoproteomics: patterning the systems biology mosaic.* Wiley Interdiscip Rev Dev Biol, 2014. **3**(1): p. 83-112.

221. Mayya, V. and D.K. Han, *Phosphoproteomics by Mass Spectrometry: insights, implications, applications, and limitations.* Expert review of proteomics, 2009. **6**(6): p. 605-618.

222. Solari, F.A., et al., *Why phosphoproteomics is still a challenge.* Mol Biosyst, 2015. **11**(6): p. 1487-93.

223. Tuli, L. and H.W. Ressom, *LC–MS Based Detection of Differential Protein Expression.* Journal of proteomics & bioinformatics, 2009. **2**: p. 416-438.

224. van den Broek, I., et al., *Current trends in mass spectrometry of peptides and proteins: Application to veterinary and sports-doping control.* Mass Spectrom Rev, 2015. **34**(6): p. 571-94.

225. Hughes, C.S., et al., *Quantitative Profiling of Single Formalin Fixed Tumour Sections: proteomics for translational research.* Scientific Reports, 2016. **6**: p. 34949.

ACKNOWLEDGEMENTS

My first acknowledgement is to the Luxembourg Institute of Health (LIH) and University of Luxembourg for granting me the opportunity to obtain my PhD in collaboration with the Fonds National de la Recherché (FNR) responsible for the financial support.

I am grateful to the members of the thesis committee, Prof. Dr. Pierre Thibault, Prof. Dr. Iris Behrmann and Prof. Dr. Serge Haan for all the knowledge transfer and input during these four years. I am also grateful to Dr. Dobrin Nedelkov and Prof. Dr. Florian Stengel for being part of my PhD defense committee along with Prof. Haan, Prof. Behrmann and Dr. Dittmar.

Prof. Serge Haan there are no words to express my gratitude to you for all the help and sacrifice during the less happy days. Thank you for being my supervisor.

I would like to thank Prof. Dr. Bruno Domon for giving me this opportunity to challenge myself and start my PhD adventure and Dr. Gunnar Dittmar for helping me to complete it and explore the more interesting biological part of my project.

I want to thank Dr. Jan van Oostrum for all the positivity, assistance and help, it was a pleasure to learn from you.

Dr. Lesur and Dr. El Khoury, apologies for not being the perfect student and big gratitude for all your help, patience and knowledge.

Cristina and Lizianne, thank you for all your help and shared time during these years, especially in the difficult moments.

Special thanks to my office colleagues, Daniela, Miriam and Adele for teaching me about the differences in cultures, expressions, laughs and kindness. It was a great pleasure to share the office with you.

Dr. Puard and M. Bernardin, merci beaucoup for all the antibody/WB help and shared French delicacy. Also an enormous merci beaucoup to Nathalie for helping me with the genomics and transcriptomics part of my thesis.

Thank you to all Proteomics group, especially to Katriina and Elenita, for all memorable moments. It was a great pleasure sharing my glass of red wine with you.

Iskra, Ema, Mone, Kiki and Marija, if it wasn't you I would have withdrawn from everything. You are my cornerstone.

And big thank you to my family for all the support.

ANNEXES

List of communications

1. **Papers:**
    a. Lesur A., **Ancheva L.** *et al., Screening protein isoforms predictive for cancer using immuno-affinity capture and fast LC-MS in PRM mode*, roteomics Clinical Applications, 2015.

    b. Trevisiol S, Ayoub D, Lesur A, **Ancheva L**, Gallien S, Domon B, *The use of proteases complementary to trypsin to probe isoforms and modifications*, Proteomics Journal, 2015.

2. **Oral presentation:**
    a. **Ancheva L**, Mass spectrometry characterization of immuno-purified cancer related protein isoforms, PhDDays, Luxembourg, 2015.

3. **Poster presentations:**
    a. **Ancheva L**, Lesur A, van Oostrum J, Domon B, *Protein Profiling of Immuno-purified samples using Mass Spectrometry,* 9th European Summer School for Advanced Proteomics, Italy, 2015.
    b. **Ancheva L.,** Lesur A, van Oostrum J, Domon B, *Mass Spectrometric Immunoassay for characterization of Cancer Related Protein Isoforms*, PhD Days, 2014.
    c. Lesur A, **Ancheva L**, Gallien S, van Oostrum J, Domon B, *Characterization of protein isoforms predictive of cancer using immuno-affinity capture and fast LC-MS*, HUPO Annual Congress, 2014.
    d. Lesur A, **Ancheva L**, Gallien S, van Oostrum J, Domon B, *Combining Immuno Affinity Purification and Fast LC-MS to Characterize Peptide Isoforms of Diagnostic Cancer Markers,* ASMS Annual Congress, 2014.

RESEARCH ARTICLE

Screening protein isoforms predictive for cancer using immunoaffinity capture and fast LC-MS in PRM mode

Antoine Lesur[1], Lina Ancheva[1], Yeoun Jin Kim[1], Guy Berchem[2,3], Jan van Oostrum[1] and Bruno Domon[1]

[1] Luxembourg Clinical Proteomics Center (LCP), CRP-Santé, Strassen, Luxembourg
[2] Laboratory of Experimental Hemato-Oncology, CRP-Santé, Strassen, Luxembourg
[3] Centre Hospitalier Luxembourg (CHL), Strassen, Luxembourg

Received: October 15, 2014
Revised: December 19, 2014
Accepted: February 2, 2015

Purpose: We report an immunocapture strategy to extract proteins known to harbor driver mutations for a defined cancer type before the simultaneous assessment of their mutational status by MS. Such a method bypasses the sensitivity and selectivity issues encountered during the analysis of unfractionated complex biological samples.
Experimental design: Fast LC separations using short nanobore columns hyphenated with a high-resolution quadrupole-orbitrap mass spectrometer have been devised to take advantage of fast MS cycle times in conjunction with sharp chromatographic peak widths to accelerate the sample analysis throughput. Such an analytical platform is well suited to analyze simple protein mixtures obtained after immunoaffinity enrichment.
Results: After establishing the technical performance of the platform, the method was applied to the quantitative profiling of cellular Ras and EGFR protein isoforms, as well as serum amyloid A isoforms in plasma.
Conclusions and clinical relevance: Immunoaffinity purification combined with fast LC-MS detection for the detection of driver mutations in tissue and tumor biomarkers in plasma samples can assist clinicians to select an optimal therapeutic intervention for patients.

Keywords:
EGFR / Fast LC / KRas / PLASMA / PRM / SAA

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

All cancers are caused by DNA changes that affect the function of certain gene products to provide expansion capabilities to the cell. Such transformation events are caused by mutations, gene fusions, or gene amplifications resulting in a growth advantage or increased survival rate [1]. Although germline mutations can result in predisposition to heritable cancers, somatic mutations represent the majority of observed genomic alterations. Somatic mutations arise in cancers but are not found in matched normal tissues from the same patient. A single genetic change is rarely sufficient for malignant tumor development and for the formation of solid tumors multiple genetic changes are required. However, only a fraction of the alterations are responsible for the initiation and progression of tumors, and these are called "driver" mutations [2]. The remaining genetic mutations are called "passenger" mutations, lacking an apparent selective growth. Solid tumors may contain 40–100 alterations distributed over the coding regions but only about 5–15 of those are considered to be driver mutations [3]. Driver mutations in oncogenes occur frequently in a heterozygous setting, whereas effective mutations in tumor suppressor genes are typically homozygous in nature. Most targeted cancer drugs are however directed against oncoproteins making quantification of the

Correspondence: Dr. Bruno Domon, Luxembourg Clinical Proteomics Center (LCP), CRP-Santé, L-1445 Strassen, Luxembourg
Fax: +352-26970-717
Email: bdomon@lih.lu

Abbreviations: **EGFR**, epidermal growth factor receptor; **KRas**, Kirsten rat sarcoma-viral oncogene homolog; **MASI**, mutant allele specific imbalance; **MSIA**, mass spectrometric immunoassay; **NSCLC**, non-small cell lung cancer; **PRM**, parallel reaction monitoring; **SAA**, serum amyloid A; **SIL**, stable isotopically labeled; **UHPLC**, ultra-HPLC; **WT**, wild type

Colour Online: See the article online to view Figs. 1–3 in colour

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Clinical Relevance

An analytical platform for the unambiguous detection of oncogenic protein mutations and isoforms based on immunoaffinity purification and fast LC-MS detection has been developed. Highly selective and high-throughput detection methods are of importance in order to support clinicians in the selection of targeted therapeutic strategies.

relative expression levels of mutant versus wild-type protein forms of importance, especially, in light of a potential mutant allele specific imbalance (MASI) [4, 5]. Although the effects of MASI at the genomic level are well described, the consequences at the protein level are poorly understood. Studies on the differential sensitivity toward targeted inhibitors as, for example, reported for heterozygous BRAF mutations compared to homozygous BRAF mutations studied in human melanomas, suggest that an increase in mutant protein concentration in a heterozygous setting with MASI in general [6] could have relevance for a future therapeutic strategy.

For a number of diseases—including non-small cell lung cancer (NSCLC), colorectal cancer, and breast cancer—there are options to bypass empirical chemotherapy and implement a personalized approach with targeted drugs in the form of small molecules or antibodies. The concept of targeted agents was introduced with the growing understanding of the tumor at a molecular level. This strategy relies on the classification of tumors in clinically relevant subsets according to the presence of driver mutations [7,8]. Tumor profiling has started by the analysis of individual genes in a cancer subtype to make predictions concerning the sensitivity of that tumor toward a drug targeting a specific protein. Examples are the test for HER2 overexpression as an indicator for positive response to trastuzumab [9]. For NSCLC, the presence of mutations in the tyrosine kinase domain of the EGF receptor was found to be strongly correlated to the response to the kinase inhibitors gefitinib and erlotinib [10].

The current trend in pharmaceutical development toward highly targeted drugs has resulted in the encouragement by the Food and Drugs Administration to supply companion diagnostic tools. As many different mutations are found to impact therapy, it would be highly cumbersome to have individual and different diagnostic tests specific for each targeted drug candidate [11]. Although this approach is highly suitable for the selection of patient populations to be enrolled in clinical trials, the high cost and sample consumption are prohibitive to screen patients for all relevant mutations using individual assays. Instead of testing tumors for the presence of individual mutations, the focus should be on tests comprising an array of predictive markers. Especially, because the presence of driver mutation is not restricted to a particular cancer type or subtype, but can be observed in a variety of different tumors. Although MS-based technologies are capable of detecting attomole quantities of proteins, the sensitivity is usually severely compromised due to the chemical complexity of the samples. Therefore, a multiplexed immunocapture strategy to enrich those proteins known to harbor driver mutations for a defined cancer type needs to be executed before the simultaneous assessment of the mutational status of the individual proteins by mass spectrometric methods. Two feasibility studies based on an immunoaffinity approach have been reported providing support for the approach to detect and quantify nonsynonymous mutations. Both studies were limited to the detection of Kirsten rat sarcoma viral oncogene homolog (KRas) mutation, either restricted to the determination of a single KRas G12D mutation [12] or to multiple Ras mutations at amino acid position 12 [13]. Another study also reported the analysis of KRas mutations albeit using an SDS-PAGE separation of the samples prior to MS analysis [14].

We have investigated short assay times enabled by the application of ultra-HPLC (UHPLC) to achieve an efficient screening platform through the use of short LC runs with a 6-min gradient time instead of the standard 60-min gradient. A targeted acquisition method using high-resolution and accurate MS has been developed with a quadrupole-orbitrap instrument. This acquisition method, called parallel reaction monitoring (PRM), is based on the signal extraction of fragment ions from high-resolution MS/MS spectra and combines high levels of sensitivity and selectivity. The ability of measuring quantitatively a set of driver mutations at the protein level with high sensitivity opens not only avenues for patient stratification but may also provide additional evidence facilitating the verification of hypotheses concerning tumor evolution and driver mutation heterogeneity in tumor tissues that are mainly based on genomic analyses.

2 Materials and methods

2.1 Reagents

Mass spectrometric immunoassay (MSIA) disposable automation research tips with protein A/G covalently bonded (cat. no. 991PRT15) were obtained from Thermo Scientific BVBA. Monoclonal mouse antipan-Ras antibody, clone Ras 10 (cat. no. MABS195) and monoclonal mouse anti-EGFR (where EGFR is epidermal growth factor receptor), clone 528 (cat. no. MABF119) were purchased from Millipore. Monoclonal mouse anti-SAA (serum amyloid A) antibody (cat. no. ab687) was purchased from Abcam. A549 and H1975

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

cell lines were a gift from the Laboratory of Experimental Hemato-Oncology (LHCE) of the CRP-Santé in Luxembourg, the Hcc827 cell line was obtained from the ATCC, and patient and control plasma samples were obtained from the Integrated Bio Bank of Luxemburg (IBBL). Sequencing grade modified trypsin was purchased from Promega. Endo-Glu-C (staphylococcus protease V8) was obtained from Worthington. Full-length human recombinant KRas protein (cat. no. ab96817) was purchased from Abcam. All other reagents were obtained from Sigma-Aldrich.

2.2 Evaluation of the fast-LC system performance in PRM mode isolated from the sample preparation

A stock solution of quantified stable isotopically labeled (SIL) Ras peptides (C-terminal arginine, $^{13}C_6$, $^{15}N_4$, $\Delta m = 10$ u, C-terminal lysine, $^{13}C_6$, $^{15}N_2$, $\Delta m = 8$ u; AQUA quantpro, Thermo Fisher) was diluted in a constant amount of trypsinized recombinant wild-type KRas and BSA to obtain concentrations ranging from 0.013 to 50 fmol/μL. The SIL peptides were provided by the manufacturer in ACN/water buffer 5% v/v at the stock concentration of 5 pmol/μL. The concentrations were measured by the manufacturer using the amino acids analysis method and the purity was established by UHPLC-UV analysis. On arrival, peptides were aliquoted in Eppendorf Protein Lobind tubes and stored at −20°C. Each aliquot is intended for a single use to avoid freeze–thaw cycles.

The BSA was employed to create a low complexity chemical background, and to prevent unspecific adhesion of peptides. The concentrations of BSA and recombinant KRas in all samples were 5 and 10 fmol/μL, respectively. Each concentration of the dilution series was injected six times, with three injection replicates designed as calibrants, and the others as quality controls. The signals of the KRas endogenous peptides were used to normalize the signal of the isotope-labeled peptides. The area ratios of calibrants were plotted against the known concentrations of isotope-labeled peptides to calculate linear regressions with a $1/x^2$ weighting. Both concentrations of calibrants and quality controls were back-calculated using the equations generated from the calibrant dilution curves.

2.3 Cell culture

A549 lung cancer adenocarcinoma cells (KRas G12S mutation and EGFR wild-type) were grown in DMEM/F12 medium (Lonza), supplemented with 10% v/v FBS (Life Technologies) and 1% v/v penicillin/streptomycin mixture (Lonza). Cells were incubated at 37°C in 95% humidity atmosphere and 5% CO_2, and grown to confluence at around 2 million cells/sample.

H1975 lung cancer adenocarcinoma cells (KRas wild-type and EGFR L858R mutation) were routinely cultured in T-75 flasks with RPMI-1640 medium (Lonza) supplemented with 10% v/v FBS and 1% v/v penicillin/streptomycin mixture to reach confluence at about 2 million cells.

Hcc827 lung epithelial adenocarcinoma cells (KRas wild-type and EGFR E746-A750 deletion mutation) were grown in supplemented RPMI-1640 medium (same as for H1975), under the same conditions described above to reach confluence at about 2 million cells.

The confluent cells were collected from T-75 flasks by washing twice with Hank's balanced salt solution (without phenol red, calcium, and magnesium; Lonza) followed by incubation with 0.02% w/v EDTA (Lonza) at 37°C for 15 min for cell detachment. Harvested cells were transferred and centrifuged for 10 min at 200 × *g*. After aspirating the medium/EDTA mixture, the pellets were resuspended in 2 mL lysis buffer containing 130 μL protease inhibitor cocktail. The samples were either immediately processed or stored at −80°C till further analysis.

2.4 Cell lysis, immunoaffinity purification, and proteolysis

2.4.1 Ras proteins

Proteins were extracted from about 3.5 and 3 million A549 and H1975 cells, respectively, with 2 mL modified RIPA lysis buffer (50 mM Tris-HCl, 150 mM NaCl, 0.1% SDS, 0.5% sodium deoxycholate, and 1% octyl β-D-glucopyranoside), supplemented with protease inhibitor cocktail, and subjected to three freeze/thaw cycles. Protein lysates were centrifuged at 4°C at 20 000 x *g* for 30 min and the supernatant was collected for enrichment using MSIA tips with protein A/G beads. The MSIA tips were previously loaded with 5 μg of anti-pan-Ras antibody according to the instructions of the manufacturer. Extraction of the Ras proteins was performed by 999 aspiration/dispense cycles of the cell lysates solution through the MSIA tips. The tips were then washed with 300 μL of 2 M ammonium acetate buffer (pH 7.3) and thrice with 150 μL of water. Bound proteins were eluted from the MSIA tips in 50 μL of water acidified with formic acid (56 mM). Eluates were vacuum-dried and further resuspended in 30 μL of 50 mM ammonium bicarbonate buffer (pH 8.0). Disulfide bonds were reduced with 5 μL of 50 mM DTT for 45 min at 50°C, followed by the alkylation of the free thiol groups with 5 μL of 150 mM iodoacetamide for 30 min at room temperature in the dark. Then, 0.05 μg of trypsin were added per sample and incubation was performed overnight at 37°C. The activity of the trypsin was quenched by adding 0.5 μL formic acid. Finally, samples were spiked with purified and quantified isotopically labeled synthetic peptides (C-terminal arginine, $^{13}C_6$, $^{15}N_4$, $\Delta m = 10$ u, C-terminal lysine, $^{13}C_6$, $^{15}N_2$, $\Delta m = 8$ u; AQUA quantpro, Thermo Fisher) and directly analyzed by fast LC-MS in PRM mode with an on-line desalting onto the trapping column.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

2.4.2 EGF receptor

Extraction and sample preparation of EGFR from the cell lines A549, H1975, and Hcc827 similar to the protocol employed for Ras analysis. However, a commercial RIPA lysis buffer (Sigma-Aldrich) was used and the MSIA tips were loaded with 5 μg of anti-EGFR antibody directed against the extracellular part of the EGFR. Proteins were eluted from MSIA tips using an aqueous trifluoroacetic acid 0.4%, 33% ACN v/v buffer. Ammonium bicarbonate was replaced by sodium phosphate buffer (pH 7.8), and 0.4 μg of Glu-C protease was added for the overnight digestion. Samples were spiked with heavy-labeled peptides (combination of internal arginine, $^{13}C_6$, $^{15}N_4$, $\Delta m = 10$ u; internal lysine, $^{13}C_6$, $^{15}N_2$, $\Delta m = 8$ u; internal alanine, $^{13}C_3$, ^{15}N, $\Delta m = 4$ u; and internal leucine, $^{13}C_6$, ^{15}N, $\Delta m = 7$ u; AQUA Quantpro ThermoFisher), desalted onto SPE Sep-PakC18 cartridges (Waters), vacuum-dried, resuspended in 25 μL of aqueous formic acid 0.1%, and analyzed by fast LC-MS in PRM mode.

2.4.3 SAA

Four microliters of plasma from six lung cancer patients (two males, four females, median age 60, stage IV), and six healthy donors (male, median age 39), were obtained from IBBL. Informed consent forms approved by the Comité National d'Ethique de Recherche (CNER) were obtained from the patients prior to sample collection. Blood samples were processed following the standard operating protocols of IBBL to prepare aliquots of plasma. Plasma samples were diluted to 100 μL final volume with 10 mM PBS at pH 7.4. The samples were purified in duplicate using 500 aspiration/dispense cycles on MSIA protein A/G tips previously loaded with an anti-SAA antibody. The MSIA tips were washed with 10 mM PBS and twice with water. Extracted proteins were then eluted from the MSIA tips in 50 μL 33% v/v ACN in water acidified with 0.4% trifluoroacetic acid. After elution, the samples were vacuum-dried, resuspended in 30 μL 50 mM ammonium bicarbonate buffer (pH 8.0), reduced with 5 μL 50 mM DTT for 45 min at 50°C, alkylated with 5 μL 150 mM iodoacetamide for 30 min at room temperature in the dark, and digested with 0.04 μg of trypsin overnight at 37°C. Samples were spiked with heavy-labeled peptide standards (crude quality, Thermo Fisher) and directly analyzed by fast LC-MS in PRM mode with an on-line desalting onto the trapping column.

2.5 LC-MS configuration

The LC-MS system consisted of Proxeon NLC-1000 operated in a column-switching setup. The mobile phases A and B consisted of 0.1% formic acid in water and ACN, respectively. Samples were injected onto the trapping column (300 μm × 5 mm, C8 pepmap100, 5 μm) using the mobile phase A delivered at a constant pressure of 600 bars. The samples were then eluted from the trapping column onto the analytical column (150 μm id × 45 mm, syncronis C18, 1.7 μm, 100 Å) at a flow rate of 1.5 μL/min by a linear gradient starting from 2% B/98% A to 40% B/60% A v/v in 6 min. The MS analysis was performed on a Q-Exactive Plus (Thermo Scientific) mass spectrometer equipped with an EASY-spray ion source. The time-segmented PRM method was performed with a quadrupole isolation window of 2 m/z units (3 m/z for SAA peptides for co-isolation and analysis of peptide isoforms with a close mass range in a single MS event), an automatic gain control target of 1×10^6 ions, a maximum fill time of 120 ms and an orbitrap resolving power of 35 000 at 200 m/z. Acquisition time windows were manually defined with minimal overlap and in a way to limit as much as possible the number of scan events per MS cycle time. Collision energies were optimized for each precursor.

2.6 Data processing

Fragment ion chromatograms, also called PRM traces, were extracted from MS raw data, and peak areas were integrated using the Skyline package [15]. This tool extracts, from MS/MS spectra, the full signal detected within a window of twice the estimated resolution at a given m/z. For an orbitrap analyzer, the extraction window of chromatographic traces is calculated using the following formula, with a resolving power of 35 000 at m/z 200:

$$\text{Width of the extraction window} = 2 \times \frac{m/z \times \sqrt{m/z}}{35\,000 \times \sqrt{200}}$$

Fragment ions were selected for quantification according to the accuracy of the mass measurement as well as the co-elution and the similarity of the fragmentation patterns between the endogenous and the respective isotope-labeled standards. For each peptide, the ratio between the sum of the PRM trace area of the endogenous peptide and the sum of the respective area of the isotope-labeled standard were calculated. For the screening of the SAA isoforms and variants, non-quantified internal standards were employed making only a comparison between patients for the same isoform or variant possible. For EGFR and KRas, quantified isotope-labeled peptides were mixed with the samples and the endogenous peptides were quantified by isotope dilution according to the equation:

$$\text{Concentration}\left(\frac{\text{molecules}}{\text{cell}}\right) = \frac{(\text{endogeneous/isotope labeled})\text{Area ratio} \times \text{internal standard concentration} \times N_A}{\text{concentration of cell}}$$

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

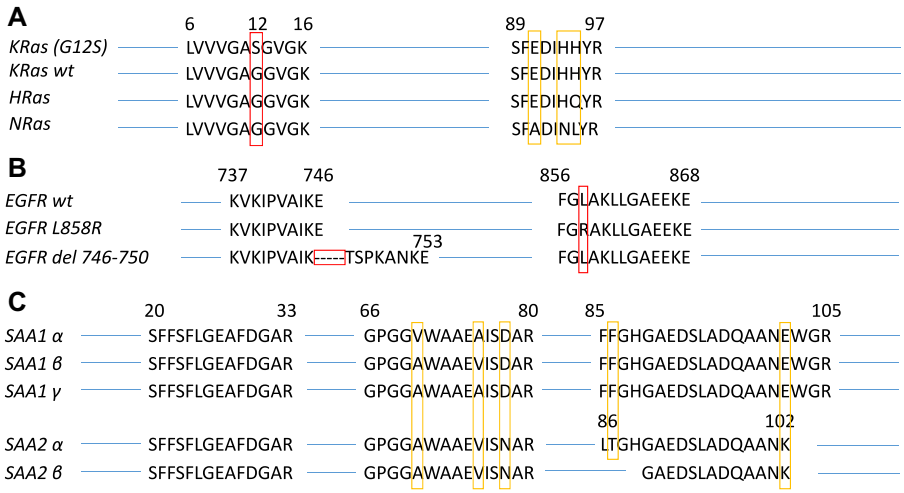


Figure 1. Sequences alignment of signature peptides of (A) Ras, (B) EGFR, and (C) SAA proteins. The positions of mutations and sequence variants are highlighted in red and yellow, respectively.

Distribution of mutants and isoforms within a cell line population was then calculated from these concentrations.

3 Results and discussion

3.1 Targeting driver mutations by fast LC-MS in PRM mode

The aim of the present work was to demonstrate first the feasibility of coupling a fast LC with a quadrupole-orbitrap mass spectrometer operated in PRM mode. Then, immunoaffinity-purified samples were analyzed on that platform for the detection of protein mutations or isoforms as diagnostic cancer markers (Supporting Information Fig. 1). EGFR, Ras, and SAA proteins (Fig. 1 and Table 1) were selected to demonstrate, as a proof of a principle, the flexibility and performance of such an analytical platform. Relative and absolute quantification, performed by isotope dilution, were investigated as well as alternatives to the classical tryptic digestion protocols.

3.2 Performance of fast LC-MS platform in PRM mode isolated from the sample preparation

An evaluation of the quantitative performance of the quadrupole-orbitrap mass spectrometer hyphenated with a fast UHPLC system was performed. UHPLC chromatography columns use sub-2 μm diameter particles as packing material for a dramatically improved separation efficiency,

Table 1. List of signature peptides and precursor charge states of endogenous and SIL

Protein	Mutation	Peptide	Precursor	
			Endogenous	Internal standard
Ras	wt	LVVVGAGGVGK	478.30 (2+)	482.31 (2+)
	G12S	LVVVGASGVGK	493.31 (2+)	497.31 (2+)
	KRas	SFEDIHHYR	401.86 (3+)	405.19 (3+)
	NRas	SFADINLYR	549.78 (2+)	554.79 (2+)
	HRas	SFEDIHQYR	398.86 (3+)	402.19 (3+)
EGFR	WT	FGLAKLLGAEE	574.32 (2+)	577.83 (2+)
	WT	FGLAKLLGAEEKE	468.93 (3+)	472.94 (3+)
	L858R	FGRAKLLGAEE	595.83 (2+)	600.83 (2+)
	L858R	FGRAKLLGAEEKE	483.27 (3+)	486.94 (3+)
	WT	KVKIPVAIKE	375.58 (3+)	378.26 (3+)
	del746-750	KVKIPVAIKTSPKANKE	463.54 (4+)	465.54 (4+)
SAA	SAA total	SFFSFLGEAFDGAR	775.87 (2+)	780.87 (2+)
	SAA 2 total	GPGGAWAAEVISNAR	728.37 (2+)	733.37 (2+)
	SAA 2 α	LTGHGAEDSLADQAANK	566.61 (3+)	569.28 (3+)
	SAA 2 β	GAEDSLADQAANK	645.30 (2+)	649.31 (2+)
	SAA 1 total	FFGHGAEDSLADQAANEWGR	726.66 (3+)	730.00 (3+)
	SAA 1 α	GPGGVWAAEAISDAR	728.86 (2+)	733.87 (2+)
	SAA 1 β	GPGGAWAAEVISDAR	728.86 (2+)	733.87 (2+)
	SAA 1 γ	GPGGAWAAEAISDAR	714.85 (2+)	719.85 (2+)

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

leading to sharper peaks and better sensitivity. In this work, short nanobore UHPLC columns (150 μm id × 45 mm), packed with 1.7 μm particles were employed with rapid elution gradients of 6 min to facilitate the high-throughput analysis of large sample sets. Such a chromatographic system elutes peptides into peaks of about 2- to 3-s duration at half height. Under these conditions, the scanning speed of the mass spectrometer becomes a critical parameter as around 8 data points must be acquired across a peak profile in order to obtain a correct measurement of the peak area. A quadrupole-orbitrap mass spectrometer operated in PRM acquisition mode was used for detection, as it allows the rapid recording of high-resolution MS/MS data. Such mass spectrometer traps ions before the scanning event in the orbitrap and therefore requires an estimation of the ion flux to predict fill times long enough to achieve high sensitivity without exceeding the capacity of the trap. This process is important, as with the fast UHPLC setup the chromatographic peak width is roughly ten times sharper than with standard LC and fast variations of the ion flux can occur.

The performance evaluation of the fast LC-PRM platform was investigated using dilution curves of quantified isotope-labeled synthetic peptides added to a recombinant KRas protein tryptic digest. The Supporting Information Table 1 summarizes the linearity range observed for two KRas wild-type signature peptides. A linear relationship has been established from 0.013 to 50 fmol/μL for LVVVGAGGVGK and from 0.033 to 50 fmol/μL for SFEDIHHYR. The inaccuracy on the back-calculated concentration of calibrants and quality controls, as well as the RSDs, is below 15%. The carry-over of the LC system after an injection of 250 fmol onto the column is about 0.5% of the original signal recovered in the next blank injection and a background below the limit of quantification is restored after two to three additional blank injections for these peptides. This is an acceptable performance for a nanoscale chromatographic system, usually prone to sample carry-over. The RSD of elution times over the 60 injections of this experiment was below 0.4%. Such elution time reproducibility is a critical parameter for time-scheduled PRM experiments, as narrower monitoring windows can be used without the need of frequent adjustment of retention time windows.

3.3 Combining immunoaffinity purification with fast LC-PRM

3.3.1 Ras family proteins

The Ras protein family includes three isoforms, KRas, Harvey rat sarcoma viral oncogene homolog (HRas), and neuroblastoma RAS viral oncogene homolog (NRas), which can be discriminated by a signature peptide in positions 89–97 (Fig. 1A). Point mutations in KRas at position G12 or G13 (Supporting Information Table 2), which are detectable in the tryptic peptides 6–12, predict the poor clinical outcome of the therapies targeting EGFR [16].

Ras proteins were purified from A549 (G12S KRas mutation) and H1975 (wild-type KRas) cell lines using disposable pipette tips packed with protein A/G beads (MSIA) previously loaded with a pan-Ras antibody. The correlation between the amount of the cell lysate loaded onto the MSIA tip and the MS signal was found to be linear (Supporting Information Fig. 2). The representative fast LC-MS PRM chromatograms of the Ras signature peptides from the A549 cell line are illustrated in Fig. 2A. The immunoaffinity purification dramatically reduced the sample complexity and unambiguous PRM traces were recorded. The SIL acts as an identity confirmation tool based on the co-elution of the chromatographic profiles as well as on the similarity of fragmentation patterns.

The chromatographic peaks, of about 3 s wide (at half height), were defined by 8–16 data points, depending on the number of concurrent eluting precursors in a time window. The RSD of the area ratios between the endogenous signature peptides and their respective SIL PRM traces ranged from 11 to 30% indicating a good reproducibility between five replications of the complete sample preparation workflow, from cell lysis to the LC-PRM analysis.

As expected, the G12S signature peptide was only detected in the A549 cell line that contains the homozygous G12S Kras mutation (Table 2). The wild-type version of the peptide was also detected in A549 cells but was derived from the homologous peptide sequence shared with NRas and HRas.

The total concentration of Ras proteins estimated from either the sum of the concentration of the signature peptides (6–16) or from the sum of the three signature peptides 89–97 identical for the H1975 cell line. This is however not the case for the A549 cell line where an excess of 1.6 times of the 6–16 peptides was detected as compared to the sum of the 89–97 peptides. Although we cannot explain this result, similar observations can be found in literature. It can be observed in results published by Wang et al. [13], where the ratio of peptides 6–16 over peptides 89–97 is 1.8- and 1.4-fold higher for the cell lines SW480 and Pa08C, which display high expression levels of the mutation, whereas cell lines with low expression levels of the mutation or those having wild-type KRas display a ratio of about 1, as we observed for the H1975 cell line [13].

3.3.2 EGFR

For NSCLC, the presence of mutations in the tyrosine kinase domain of the EGF receptor was found to be strongly correlated to the response to the kinase inhibitors gefitinib and erlotinib [10]. Two EGFR mutations account for the vast majority of sensitizing mutations in lung cancer. EGFR exon 19 deletions are in-frame deletions occurring within exon 19 that encodes part of the kinase domain. This mutation occurs with a frequency of approximately 48% in EGFR mutant lung tumors [17]. The L858R mutation results in an amino acid substitution at position 858 in EGFR. This mutation occurs within exon 21 that encodes part of the kinase domain,

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

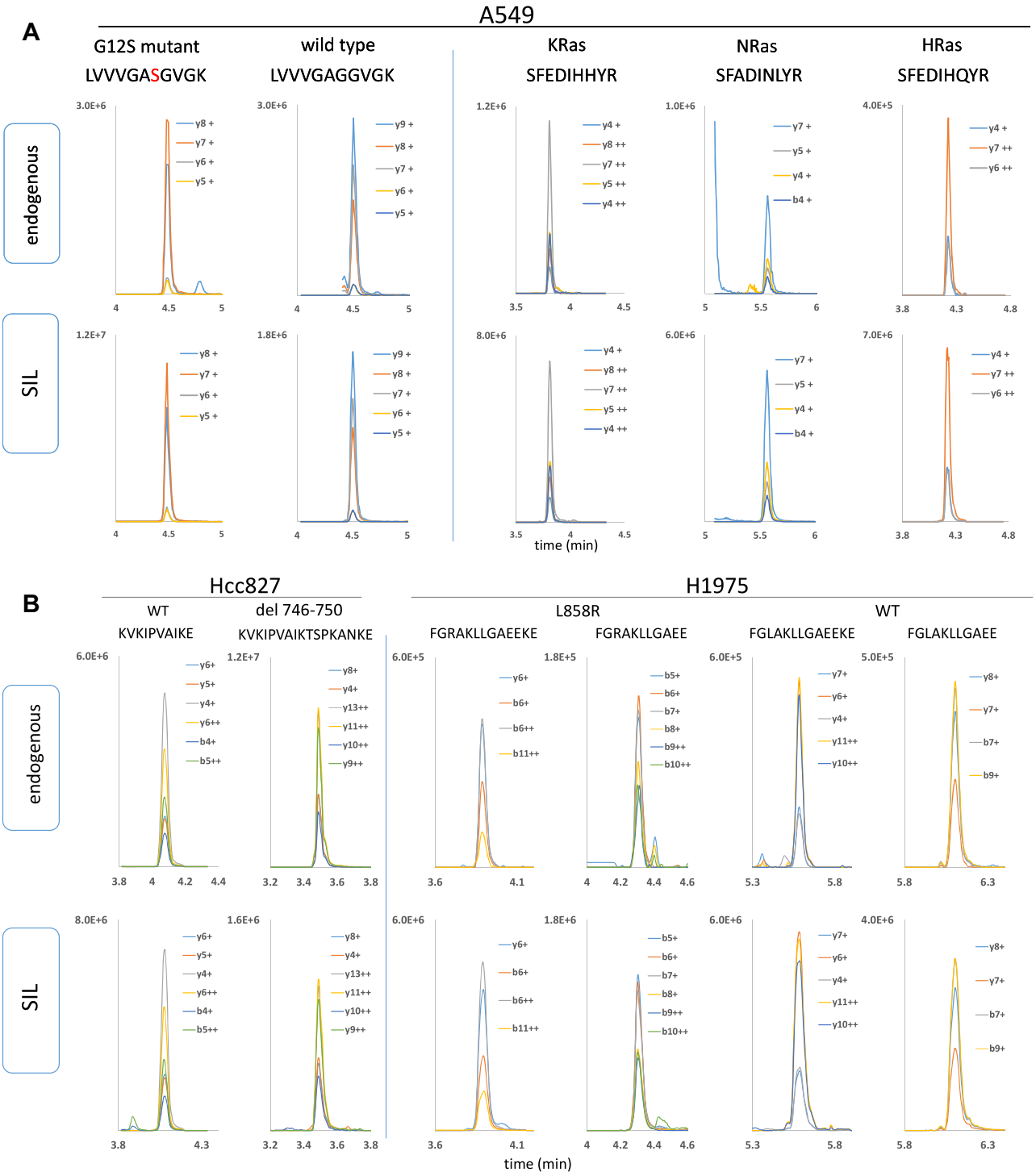


Figure 2. PRM chromatograms of (A) Ras isoforms and mutant signature peptides detected in A549 cells, the equivalent of 3e5 cells has been injected; (B) EGFR mutant signature peptides detected in Hcc827 and H1975 cells, the equivalent of 2e5 cells has been injected.

and occurs with a frequency of approximately 43% in EGFR mutant lung tumors [17].

MS-based proteomics is typically an indirect approach, as peptides generated by proteolytic digestion are measured as surrogate of proteins. Trypsin is the most favored protease because it yields high digestion efficiency and generates peptides with clear MS/MS fragmentation spectra. However, to target localized changes in the sequence of a mutated protein, the surrogate peptide must cover the mutated part of the sequence. For some cases, a tryptic peptide can be either too

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 2. Quantitative distribution of mutation G12S and KRas, NRas, and HRas isoforms for the cell lines H1975 and A549, and del 746–750 and L858R mutation of EGFR for the cell lines H1975, A549, and Hcc827

			H1975 (%)	A549 (%)	Hcc827 (%)
Ras	G12S	Mutant	0	66	-
		wt	100	34	-
	Isoforms	KRas	31	50	-
		NRas	58	34	-
		HRas	11	17	-
EGFR	del 746–750	Mutant	0	0	90
		wt	100	100	10
	L858R	Mutant	50	0	0
		wt	50	100	100

short or too long and thus cannot be properly separated by reverse phase chromatography and fragmented under low-energy CID. In such conditions, the alternative digestion patterns of other proteases, including Glu-C, Asp-N, Lys-C, Arg-C, and chymotrypsin can be investigated to produce peptides with sufficient length and overall MS compatibility [18]. Although for example, nonsynonymous mutations in KRas at position amino acid 12 or 13 can be investigated after tryptic digestion, such approach for EGFR will not result in suitable peptides for MS analysis to describe the activating mutations based on exon 19 deletions or the nonsynonymous mutation L858R. For instance, the signature peptide of the deletion 746–750 mutant of EGFR contains five lysine residues and this portion of the protein sequence is not covered by a classical trypsin digestion. This problem can be circumvented using Glu-C protease that cleaves after glutamic (E) and aspartic (D) acids, and results in peptides suitable for MS analysis (Fig. 1B).

Using a similar workflow as for Ras but with trypsin substituted by Glu-C protease, three cell lines including A549 (wt), Hcc827 (del 746–750), and H1975 (L848R) were investigated and the representative PRM chromatograms of the signature peptides detected in cell lines Hcc827 and H1975 are presented in Fig. 2B. This experiment illustrates that Glu-C peptides, despite carrying a significant number of charged amino acids, can be properly separated by chromatography and detected and quantified without ambiguity by MS. The RSD of the area ratios between the endogenous signature peptides and their respective isotope-labeled PRM traces ranged from 6 to 23% between five replications of the complete sample preparation procedure. However, the Glu-C protease seems prone to miss cleavages; the protease has been described not to efficiently cleave an EE peptide bond [19], which is the reason why the peptide FG(L/R)AKLLGAE is not detectable. However, similar intensities for the peptides FG(L/R)AKLLGAEE and FG(L/R)AKLLGAEEKE have been observed and both peptides were monitored. The ratio of the total concentration of the signature peptides, that cover the deletion mutation over the total concentration of signature peptides covering the point mutation, is about 2.2 with an RSD of 6% between the three cell lines. This indicates the good consistency of the Glu-C proteolysis.

As expected, the exon 19 deletion mutation and the L858R point mutation were detected in cell lines Hcc827 and H1975, respectively. H1975 is heterozygous for the EGFR point mutation, and an equal distribution of mutated and wild-type peptides was observed in this cell line (Table 2). However, although in Hcc827 the EGFR exon19 deletion is also heterozygous, the EGFR genes have been in addition reported to be amplified resulting in an increased sensitivity to EGFR inhibitors [20]. This increased sensitivity toward kinase inhibitors can be attributed to the observed ninefold excess of deletion mutant over wild-type EGFR in the Hcc827 cell line. In addition the total EGFR concentration, calculated by either the sum of KVKIPVAIKE and KVKIPVAIKTSPKANKE peptides or by the sum of FG(L/R)AKLLGAEE/KE peptides, was found to be 8 and 11 times higher (7% RSD) in Hcc827 cell line as compared to the H1975 and A549 cell lines, respectively. These results correlate with the observations of an overexpression of EGFR in Hcc827 cell [20].

3.3.3 Screening of SAA isoforms in human plasma

SAA1 and SAA2 belong to the acute-phase proteins secreted from the liver in response to infections and tissue injury [21]. The newly discovered roles for SAA in innate immunity during cancer progression [22,23] and in metastatic pathogenesis of lung cancer [24] raised renewed interest. High levels of expression of SAA protein in plasma can be related to chronic inflammatory states including lung cancer adenocarcinoma. SAA1 and SAA2 are highly homologous and eight signature peptides quantify the target proteins at three different levels: (Fig. 1C) variant-specific (SAA 1α, SAA 1β, SAA 1γ, SAA 2α, SAA 2β), protein-specific (SAA1 or SAA2), and pan-SAA (SAA1 and SAA2) have been used to differentiate SAA1 and SAA2, and their variants in lung cancer plasma samples as described in a previous study using a standard LC-MS approach (Kim, Y.J. et al., 2014, Specific quantification of SAA1 and SAA2 isotypes in human plasma using parallel reaction monitoring, submitted for publication.

The SAA isoforms were purified by immunoaffinity using 4 μL of plasma from six controls and six lung cancer patients (stage IV), and the complete sample preparation was replicated in duplicates. The RSD of the peptides measurement between replicates ranges between 1 and 29%, indicating a similar reproducibility as observed for Ras protein. All peptides were fully discriminated by the fast LC-MS in PRM mode with the exception of GPGGAWAAEVISDAR that had a shoulder in the front due to GPGGAWAAEVISNAR (Fig. 3A); this portion of the signal was excluded for the area integration.

In our small sample set, total SAA and SAA1 isoform levels do not seem to be specific as a marker for adenocarcinoma as high levels can be found in both patient and control samples

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

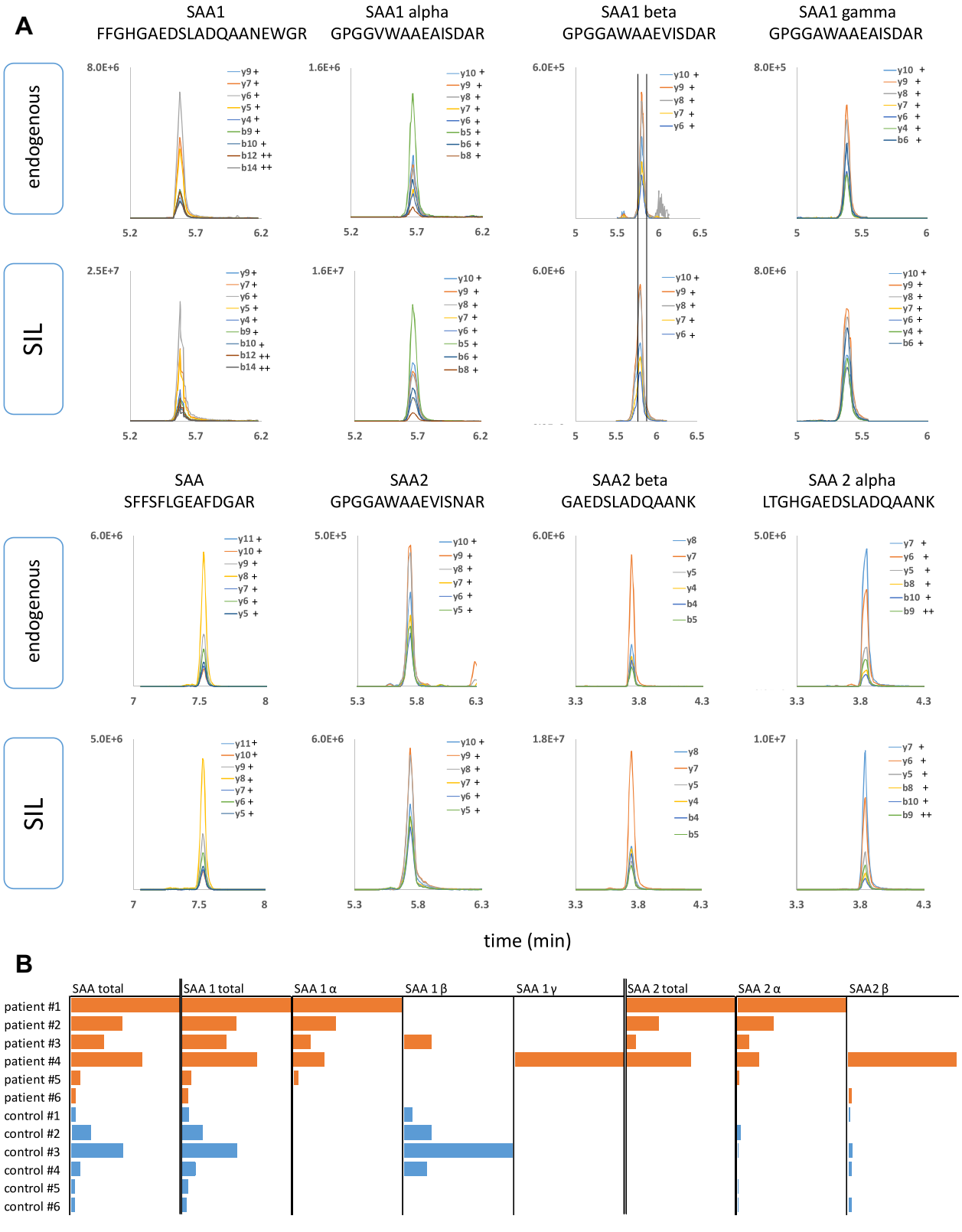


Figure 3. (A) PRM chromatograms of SAA protein isoforms extracted from human plasma. (B) Quantitative profiling of signature peptides of SAA isoforms in control and lung cancer patients plasma. Bars represent the endogenous peptide/SIL area ratios normalized to the highest value for each isoforms.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

(Fig. 3B). The ratios of SAA2α/SAA however are observed to be higher in patients regardless of the total concentration of SAA. Screening a much larger set of samples will however be required to investigate this observation further. The fast LC-MS method described in this paper would be very well suited for such a screening task. It is of interest to note that a single Caucasian patient was typed with the SAA1 gamma variant. The occurrence of this variant is common in the Asian population, but has a very low incidence (5.3%) in the Caucasian population [25].

4 Concluding remarks

Antibody-based approaches for protein detection in tissues comprise immunohistochemistry or RP protein array applications. However, mutation-specific antibodies are very rare, making immunohistochemistry or RP protein array approaches unsuitable for the detection of cancer-specific driver mutations in tumor tissues. ELISA (and variations hereof) is the current standard in protein analytics for clinical applications and is based on the immunoaffinity recognition and subsequent capture of the target protein followed by a quantitative detection methodology. Such immunoassay approaches typically do not differentiate the modifications found at the protein level, such as mutations, insertions, and deletions. MSIA was primarily developed to enable the detection and identification of peptides and proteins in biological fluids [26], acting as an alternative to ELISA-based methods and also relying on immobilized antibodies to isolate analytes from a biological sample with a subsequent release of the analytes and an MS-based analysis. MS assays in the form of SRM can typically measure in a single analysis number of preselected ions, known a priori to represent those peptides that can contain driver mutations (amino acid changes, deletions, or/and insertions). Recently, PRM, leveraging new high-resolution and accurate mass spectrometers has been introduced and high-sensitivity assays can be achieved in combination with an immunoaffinity purification [27]. PRM is the method of choice for the selective detection of specific driver mutation as it fully records the MS/MS spectra, allowing for unambiguous localization of the mutations. As the chemical background of the sample is dramatically reduced after an immunoaffinity purification, the need of a long gradient chromatography becomes obsolete and a fast LC becomes more advantageous. This combination fits better the routine screening of cancer mutations in a clinical setting. The analysis can be performed from either cancer tissues obtained from resections or biopsies, for example in the form of fine-needle aspirates, or from bodily fluids as exemplified for the SAA isoforms. The fast LC-MS platform in PRM mode can perform about 100 analyses/24 h with an overall sample preparation of about 24–36 h including an overnight proteolysis, providing a quick result for clinicians in the process of decision making of therapeutic strategy based on the mutation status. The sample preparation throughput can be further increased by multiplexed approaches using the 96-well plate format and a fast digestion procedure [28]. In addition, the method can be valuable in the stratification of patients before the start of clinical trials for targeted therapies in an oncological setting and other disease indications.

The present work has been funded by the Fonds National de la Recherche du Luxembourg (FNR) via the PEARL-CPIL program to BD and an AFR grant to AL. Authors acknowledge the Integrated Bio Bank of Luxembourg (IBBL) for collection and handing of clinical samples.

The authors have declared no conflict of interest.

5 References

[1] Croce, C. M., Oncogenes and cancer. *N. Engl. J. Med.* 2008, *358*, 502–511.

[2] Stratton, M. R., Campbell, P. J., Futreal, P. A., The cancer genome. *Nature* 2009, *458*, 719–724.

[3] Bozic, I., Antal, T., Ohtsuki, H., Carter, H. et al., Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl. Acad. Sci. USA* 2010, *107*, 18545–18550.

[4] Chiosea, S. I., Sherer, C. K., Jelic, T., Dacic, S., KRAS mutant allele-specific imbalance in lung adenocarcinoma. *Modern Pathol.* 2011, *24*, 1571–1577.

[5] Soh, J., Okumura, N., Lockwood, W. W., Yamamoto, H. et al., Oncogene mutations, copy number gains and mutant allele specific imbalance (MASI) frequently occur together in tumor cells. *PloS One* 2009, *4*, e7464.

[6] Sondergaard, J. N., Nazarian, R., Wang, Q., Guo, D. et al., Differential sensitivity of melanoma cell lines with BRAFV600E mutation to the specific Raf inhibitor PLX4032. *J. Transl. Med.* 2010, *8*, 39.

[7] Brambilla, E., Gazdar, A., Pathogenesis of lung cancer signalling pathways: roadmap for therapies. *Eur. Respir. J.* 2009, *33*, 1485–1497.

[8] Pao, W., Girard, N., New driver mutations in non-small-cell lung cancer. *Lancet Oncol.* 2011, *12*, 175–180.

[9] Vogel, C. L., Cobleigh, M. A., Tripathy, D., Gutheil, J. C. et al., Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. *J. Clin. Oncol.* 2002, *20*, 719–726.

[10] Pao, W., Miller, V., Zakowski, M., Doherty, J. et al., EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proc. Natl. Acad. Sci. USA* 2004, *101*, 13306–13311.

[11] Levy, M. A., Lovly, C. M., Pao, W., Translating genomic information into clinical medicine: lung cancer as a paradigm. *Genome Res.* 2012, *22*, 2101–2108.

[12] Ruppen-Canas, I., Lopez-Casas, P. P., Garcia, F., Ximenez-Embun, P. et al., An improved quantitative mass spectrometry analysis of tumor specific mutant proteins at high sensitivity. *Proteomics* 2012, *12*, 1319–1327.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

[13] Wang, Q., Chaerkady, R., Wu, J., Hwang, H. J. et al., Mutant proteins as cancer-specific biomarkers. *Proc. Natl. Acad. Sci. USA* 2011, *108*, 2444–2449.

[14] Halvey, P. J., Ferrone, C. R., Liebler, D. C., GeLC-MRM quantitation of mutant KRAS oncoprotein in complex biological samples. *J. Proteome Res.* 2012, *11*, 3908–3913.

[15] Sherrod, S. D., Myers, M. V., Li, M., Myers, J. S. et al., Label-free quantitation of protein modifications by pseudo selected reaction monitoring with internal reference peptides. *J. Proteome Res.* 2012, *11*, 3467–3479.

[16] Suda, K., Tomizawa, K., Mitsudomi, T., Biological and clinical significance of KRAS mutations in lung cancer: an oncogenic driver that contrasts with EGFR mutation. *Cancer Metastasis Rev.* 2010, *29*, 49–60.

[17] Mitsudomi, T., Yatabe, Y., Epidermal growth factor receptor in relation to tumor development: EGFR gene and cancer. *FEBS J.* 2010, *277*, 301–308.

[18] 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*, 1573–1584.

[19] 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*, 2833–2844.

[20] Engelman, J. A., Mukohara, T., Zejnullahu, K., Lifshits, E. et al., Allelic dilution obscures detection of a biologically significant resistance mutation in EGFR-amplified lung cancer. *J. Clin. Investig.* 2006, *116*, 2695–2706.

[21] Paret, C., Schon, Z., Szponar, A., Kovacs, G., Inflammatory protein serum amyloid A1 marks a subset of conventional renal cell carcinomas with fatal outcome. *Eur. Urol.* 2010, *57*, 859–866.

[22] Mattarollo, S. R., Smyth, M. J., A novel axis of innate immunity in cancer. *Nat. Immunol.* 2010, *11*, 981–982.

[23] Indovina, P., Marcelli, E., Maranta, P., Tarro, G., Lung cancer proteomics: recent advances in biomarker discovery. *Int. J. Proteomics* 2011, *2011*, 726–869.

[24] Sung, H. J., Ahn, J. M., Yoon, Y. H., Rhim, T. Y. et al., 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*, 1383–1395.

[25] Booth, D. R., Booth, S. E., Gillmore, J. D., Hawkins, P. N., Pepys, M. B., SAA1 alleles as risk factors in reactive systemic AA amyloidosis. *Amyloid* 1998, *5*, 262–265.

[26] Nelson, R. W., Krone, J. R., Bieber, A. L., Williams, P., Mass spectrometric immunoassay. *Anal. Chem.* 1995, *67*, 1153–1158.

[27] Peterman, S., Niederkofler, E. E., Phillips, D. A., Krastins, B. et al., An automated, high-throughput method for targeted quantification of intact insulin and its therapeutic analogs in human serum or plasma coupling mass spectrometric immunoassay with high resolution and accurate mass detection (MSIA-HR/AM). *Proteomics* 2014, *14*, 1445–1456.

[28] Lesur, A., Varesio, E., Hopfgartner, G., Accelerated tryptic digestion for the analysis of biopharmaceutical monoclonal antibodies in plasma by liquid chromatography with tandem mass spectrometric detection. *J. Chromatogr. A* 2010, *1217*, 57–64.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Supplementary material

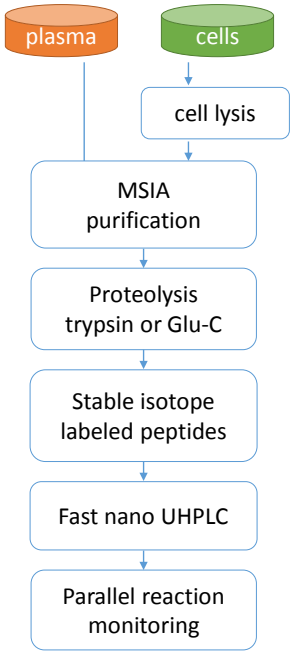
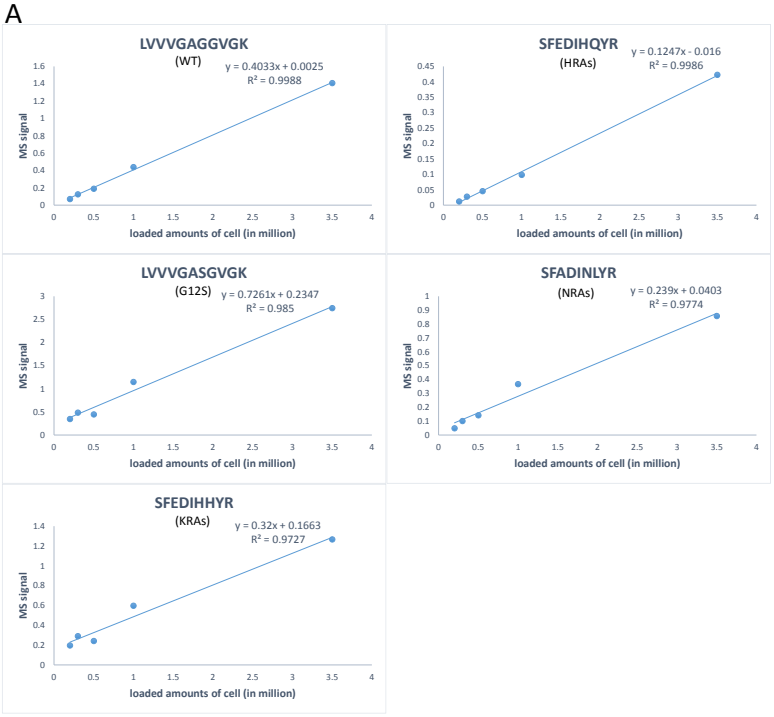
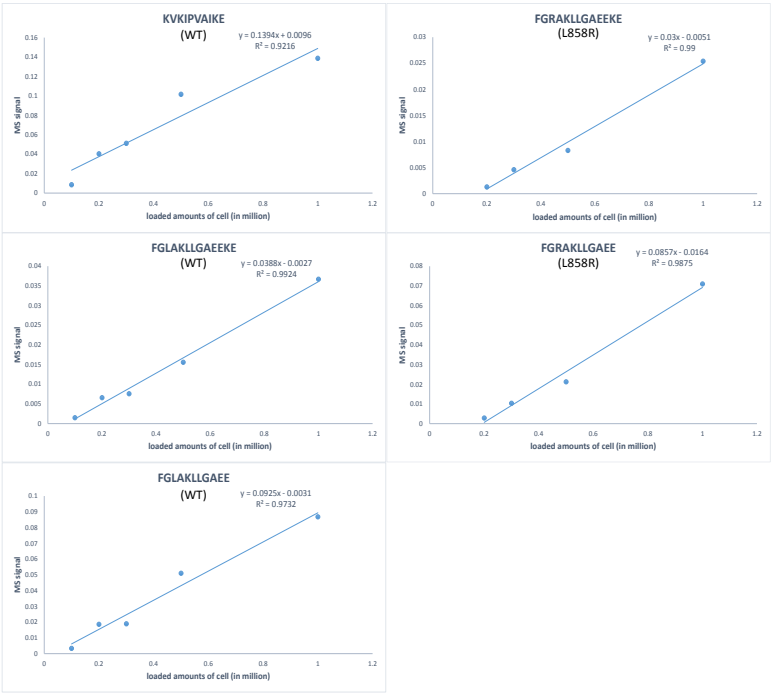


Figure S1 *Analytical workflow*

A


LVVVGAGGVGK
(WT)
y = 0.4033x + 0.0025
R² = 0.9988
SFEDIHQYR
(HRAs)
y = 0.1247x - 0.016
R² = 0.9986
LVVVGASGVGK
(G12S)
y = 0.7261x + 0.2347
R² = 0.985
SFADINLYR
(NRAs)
y = 0.239x + 0.0403
R² = 0.9774
SFEDIHHYR
(KRAs)
y = 0.32x + 0.1663
R² = 0.9727
MS signal
loaded amounts of cell (in million)

B


KVKIPVAIKE
(WT)
y = 0.1394x + 0.0096
R² = 0.9216
FGRAKLLGAEEKE
(L858R)
y = 0.03x - 0.0051
R² = 0.99
FGLAKLLGAEEKE
(WT)
y = 0.0388x - 0.0027
R² = 0.9924
FGRAKLLGAEE
(L858R)
y = 0.0857x - 0.0164
R² = 0.9875
FGLAKLLGAEE
(WT)
y = 0.0925x - 0.0031
R² = 0.9732
MS signal
loaded amounts of cell (in million)

C

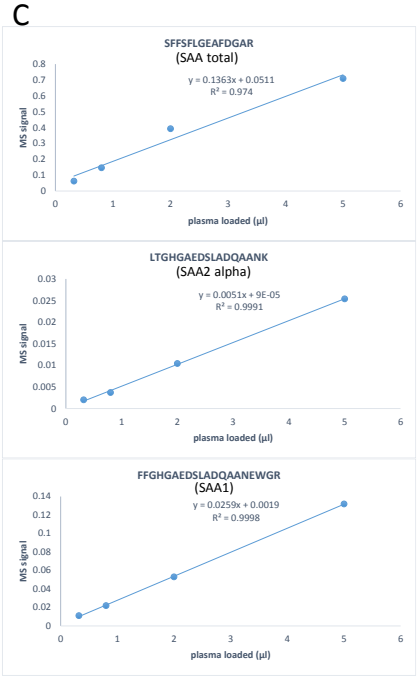


Fig S2 *Correlation between the injected amount of sample on MSIA immuno-affinity column and the recovered MS signal after elution and protease digestion.*

A) Ras protein, cell line A549

B) EGFR cell line A549 (WT) and H1975 (L858R)

C) Pool of plasma leftover

Table S1 *Quantitative performance of the Q-Exactive Plus hyphenated with the fast LC system. The concentrations of quality controls injections are back calculated using the linear regression of corresponding calibrants curves. The linear regression is weighted ($1/x^2$)*

Accuracy (%) = 100× (measured concentration / theoretical concentration)

LVVVGAGGVGK		**calculated concentration**				**accuracy**	**precision**
	femtomole/µl	**std#1**	**std#2**	**std#3**	**mean**	**mean**	**RSD %**
standard	**50.00**	52.88	46.40	51.99	50.42	101%	7%
	20.00	22.28	23.94	22.53	22.92	115%	4%
	8.00	8.55	9.12	8.63	8.77	110%	4%
	3.20	3.01	2.93	3.11	3.02	94%	3%
	1.28	1.26	1.11	1.22	1.20	94%	7%
	0.512	0.487	0.548	0.517	0.517	101%	6%
	0.205	0.192	0.199	0.190	0.194	95%	2%
	0.082	0.082	0.077	0.076	0.078	96%	4%
	0.033	0.029	0.031	0.030	0.030	92%	3%
	0.013	0.014	0.013	0.014	0.014	104%	1%
quality control	**50.00**	49.04	52.59	52.04	51.41	103%	4%
	20.00	21.53	22.52	22.99	22.35	112%	3%
	8.00	7.92	8.20	8.72	8.28	103%	5%
	3.20	2.84	3.24	3.02	3.03	95%	7%
	1.28	1.22	1.15	1.32	1.23	96%	7%
	0.512	0.465	0.524	0.487	0.492	96%	6%
	0.205	0.190	0.194	0.186	0.190	93%	2%
	0.082	0.086	0.075	0.077	0.080	97%	7%
	0.033	0.033	0.032	0.032	0.032	99%	2%
	0.013	0.012	0.014	0.012	0.013	98%	7%
slope		0.249	0.242	0.245			
intercept		-0.00079	0.00004	-0.00024			
R^2 (weighted)		0.99	0.99	0.99			

SFEDIHHYR		calculated concentration				accuracy	precision
	femtomole/µl	std#1	std#2	std#3	mean	mean	RSD %
standard	**50.00**	45.03	41.93	43.07	43.34	87%	4%
	20.00	20.20	19.42	20.16	19.93	100%	2%
	8.00	8.32	8.08	7.72	8.04	101%	4%
	3.20	3.25	3.47	3.35	3.35	105%	3%
	1.28	1.34	1.33	1.38	1.35	106%	2%
	0.512	0.533	0.554	0.536	0.541	106%	2%
	0.205	0.198	0.205	0.212	0.205	100%	4%
	0.082	0.079	0.079	0.078	0.079	96%	1%
	0.033	0.034	0.033	0.033	0.033	102%	0%
quality control	**50.00**	46.05	46.47	42.87	45.13	90%	4%
	20.00	19.12	19.62	18.55	19.10	95%	3%
	8.00	7.42	7.94	7.67	7.68	96%	3%
	3.20	3.33	3.43	3.39	3.39	106%	2%
	1.28	1.35	1.38	1.35	1.36	106%	2%
	0.512	0.512	0.564	0.550	0.542	106%	5%
	0.205	0.215	0.213	0.212	0.213	104%	1%
	0.082	0.077	0.084	0.078	0.080	97%	5%
	0.033	0.036	0.038	0.032	0.035	108%	8%
slope		0.343	0.334	0.339			
intercept		0.001	0.004	0.006			
R^2 (weighted)		1.00	0.99	0.99			

Table S2 *Frequently occurring KRas mutations at G12 and G13 positions*

			precursor	
protein	mutation	peptide	endogenous	internal standard
Ras	G13D	LVVVGAGDVGK	507.30 (2+)	511.31 (2+)
	G12S	LVVVGASGVGK	493.31 (2+)	497.31 (2+)
	G13S	LVVVGAGSVGK	493.31 (2+)	497.31 (2+)
	G13A	LVVVGAGAVGK	485.31 (2+)	489.32 (2+)
	G12A	LVVVGAAGVGK	485.31 (2+)	489.32 (2+)
	G13C	LVVVGAGCVGK	529.81 (2+)	533.81 (2+)
	G12C	LVVVGACGVGK	529.81 (2+)	533.81 (2+)
	G12V	LVVVGAVGVGK	499.32 (2+)	503.33 (2+)
	G12D	LVVVGADGVGK	507.30 (2+)	511.31 (2+)

REVIEW

The use of proteases complementary to trypsin to probe isoforms and modifications

*Stéphane Trevisiol**, *Daniel Ayoub**, *Antoine Lesur, Lina Ancheva, Sébastien Gallien and Bruno Domon*

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

Received: September 17, 2015
Revised: November 6, 2015
Accepted: December 8, 2015

The wide diversity of proteins expressed in a cell or a tissue as a result of gene variants, RNA editing or PTMs results in several hundred thousand distinct functional proteins called proteoforms. The large-scale analysis of proteomes has been driven by bottom-up MS approaches. This allowed to identify and quantify large numbers of gene products and perform PTM profiling which yielded a significant number of biological discoveries. Trypsin is the gold standard enzyme for the production of peptides in bottom-up approaches. Several investigators argued recently that the near exclusive use of trypsin provided only a partial view of the proteome and hampered the discovery of new isoforms. The use of multiple proteases in a complementary fashion can increase sequence coverage providing more extensive PTM and sequence variant profiling. Here the various approaches to characterize proteoforms are discussed, including the use of alternative enzymes to trypsin in shotgun approaches to expand the observable sequence space by LC-MS/MS. The technical considerations associated with the use of alternative enzymes are discussed.

Keywords:
Alternative proteases / Mass spectrometry / Proteoforms / Sequence coverage / Technology / Trypsin

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

At the start of the human genome project in the late 90s, the number of human coding genes was estimated to be around 100 000 [1]. That number has shrunk to around 20 000 due to the refinement of predicting tools [2]. A gene, comprised of introns and protein coding exons, is transcribed to precursor mRNAs (pre-mRNA) that undergo processing in the spliceosome which joins the exons to generate a translatable transcript (mRNA). High-throughput sequencing studies suggest that around 95% of human pre-mRNA sequences containing multiple exons yield splice variant mRNAs [3, 4]. Therefore a single gene produces several transcripts, hence multiple protein sequences. The most abundant form is called the canonical form, while the alternatively spliced variants are called isoforms. Recently, the term "proteoform" was introduced to designate the distinct protein forms resulting from post-transcriptional and post-translational modifications [5–7]. The number of alternative splicing events occurring in humans is estimated at 100 000 [8]. The extent of alternatively spliced mRNAs that undergo translation into a protein is yet unknown [6]. Despite the many cases in which discernible functional differences between splice variants have not been observed [8, 9], several studies proved that alternative splicing can produce functionally distinct protein isoforms [8, 10–17]. Splicing variants are therefore a major source of protein diversity. Other factors also contribute to the expansion of protein sequence diversity, including alternative transcription start sites and alternative polyadenylation sites, yielding shorter or longer proteins; and single nucleotide polymorphisms (SNP) which result in single amino-acid

Correspondence: Prof. Dr. Bruno Domon, Luxembourg Clinical Proteomics Center (LCP), Luxembourg Institute of Health, Strassen, Luxembourg
E-mail: bruno.domon@lih.lu
Fax: +352-26970-717

*Both authors contributed equally to this work.
****Colour Online**: See the article online to view Fig. 7 in colour.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

mutations. The resulting isoforms are protein backbone sequence variants; they vary in length sharing common exons, including variable exons, and amino-acid mutations. All this combined with all possible PTMs expands the proteome to hundreds of thousands of distinct proteoforms. It is the proteoforms that are the primary regulators of the fate and functions of the cells [18] mainly due to the PTMs that occur on specific amino acids during or after translation. A large number of possible PTMs remain largely unexplored. The most common ones include phosphorylation (S, T, Y), glycosylation (N, S, T), N-terminal acetylation, methylation (K, R), formylation (K), ubiquitinylation (K) or SUMOylation (K). The identification and characterization of PTMs in large-scale studies often require enrichment techniques and particular experimental settings rendering their study challenging.

During the last two decades, MS has driven the advancement of proteomics [19]. From an MS-centric point of view, protein sequencing is performed through two distinct approaches, top-down and bottom-up [20]. Top-down approaches consist in the MS/MS analysis of intact proteins while bottom-up consist in the analysis of peptides generated by chemical or enzymatic cleavage of proteins. Intrinsically, top-down would be the method of choice to analyze proteoforms in their integrity. It allows for the identification of the combinatorial pattern of different PTMs spread across the protein backbone as well as sequence modifications. The main challenges of top-down approaches are protein separation prior to MS analysis and MS^2 fragmentation. Top-down has been limited for a long time to the analysis of single proteins or very simple mixtures [21]. Fragmenting large proteins and achieving comprehensive sequence coverage is also challenging. Electron transfer/capture dissociation (ETD/ECD), collision induced dissociation (CID) as well as photon induced dissociation (IR or UV) have been employed [22–26]. Recent developments in ETD contributed greatly to the effectiveness of fragmentation of large proteins [27]. The combination of two or more fragmentation techniques can be used to increase fragmentation efficiency and generate more product ions hence increasing sequence coverage [26, 28]. Top-down MS for large proteins has been mainly applied to the characterization of purified recombinant proteins such as 150 kDa monoclonal antibodies. It allowed for the assignment and the identification of a significant number of PTMs including glycosylation, C-terminal lysine clipping and N-terminal glutamine cyclization [25, 29, 30]. Protein separation techniques have known a fast development in recent years allowing top-down analysis of more complex biological samples including whole cell lysates [21]. More than 5000 proteoforms corresponding to 1220 unique proteins have been identified using a combination of separation techniques and top-down MS [31]. Recently, a consortium was established to promote this approach and share the data [32]. It is clear that this approach under development is very promising and will provide major inputs to the understanding of protein and proteoform functions. Several reviews have discussed and described top-down approaches [18, 20, 21, 33, 34].

As LC-MS/MS analysis of peptides is straightforward (compared to that of proteins), bottom-up approaches allow for high-throughput proteome analysis and the identification of several thousands of proteins in a single experiment [35, 36]. However, the digestion of proteins into peptides sacrifices the integrity of proteoforms and hence reconstituting the different entities is very challenging. Nevertheless, bottom-up proteomics is by far the most widely used technique for the identification, characterization and quantification of proteins in both simple and complex mixtures. The vast majority of protein sequences in databases with expression evidence at the protein level as well as PTM sites were identified by bottom-up. Furthermore, the vast majority of data uploaded in public proteomics data repositories were acquired through this approach. Bottom-up MS relies largely on trypsin digestion and protein databases for the identification of peptides, and hence proteins; but this reliance does present drawbacks.

The protein databases are by far not exhaustive and do not contain all possible splice variants and other sequence isoforms. The high-quality database for human proteins, NeXtprot, comprises 20 060 canonical sequences among which only 10 520 are presented with one or more alternative sequences, mainly splice variants. Therefore 9540 entries contain only a canonical sequence. The total number of sequences reported (canonical plus alternative) is currently 41 980. The actual number of isoforms is much larger, especially if SNPs and other variants are taken into account. The proteogenomic approaches, in which MS spectra of peptides are mapped against protein coding genes, are better suited for the discovery of new sequence variants. Gene - expressed sequence tags (EST), and transcript databases are used to map peptides to gene or RNA sequences [37]. Blakeley et al. used the coding DNA sequences/exon genomic coordinates from the Ensembl database and mapped peptides to them. Intron spanning peptides, i.e. peptides that are on the junction of two exons confirm the exonic order. Peptides can be shared by all or a part of possible isoforms or be specific to certain isoforms. These authors suggested using targeted approaches to analyze isoform-specific peptides and unambiguously identify isoforms. They also proposed using multiple databases to account for the different annotation qualities and coverage of alternative splicing events [38]. Another elegant approach is to collect RNA- sequences and proteomics data from the same cell population. Sheykmann et al. developed a bioinformatics pipeline that detects splice junctions and translates them to amino acid sequences to be used for proteomics data search. This allowed the identification of 57 new splice junction peptides not present in the Uniprot/TrEMBL database [6]. Several recent publications have reviewed the proteogenomic approaches [39–41].

Trypsin provides peptides suitable for MS that allow the highest sequence coverage and the largest number of identifications in a complex biological sample making it the gold standard protease for LC-MS/MS analysis of digests. However, some parts of protein sequences are not accessible to trypsin due to an uneven distribution of their cleavage sites

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

(lysine and arginine). These missing sequences can contain important information such as PTMs, mutations, etc. Other enzymes have been described and were used to supplement the trypsin protein sequences. Several studies and reviews argue correctly that this dominance of trypsin has a negative outcome on proteomics analysis as it only provides a partial view of the proteome as stated by Heck and colleagues [42], leaving behind valuable information that would help identify new proteins and isoforms, more PTMs and of course achieve better sequence coverage [38, 42–47]. Furthermore, alternative digestions would provide proteotypic peptides to distinguish splice variants or other sequence variants. All these studies plead in favor of using alternative enzymes in addition to trypsin to address these issues and increase the proteome coverage. 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 cell proteome. To estimate the total sequence coverage and digestion complementarity, they measured the "proteome amino acid coverage" (PAAC). While the combination of multiple digestions did not increase significantly the number of protein groups identified, it increased the PAAC by threefold 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 [43].

A discussion on the input of complementary enzymes and an assessment of their effect on the experimental workflow of the analysis is presented here.

2 Bottom-up proteomics approaches

Two major bottom-up approaches emerged throughout the development of MS-based proteomics: discovery and targeted approaches [19]. Discovery proteomics experiments are commonly carried out using a shotgun method, which is based on data-dependent acquisition (DDA) [48, 49]. It has long been employed in the early stage of biomarker discovery including comparative studies. It allows the identification and the quantification of a large number of proteins (up to 100 00 in recent studies [35, 36]) in complex biological samples. However, the heuristic nature of precursor ion sampling affects reproducibility [50–52] and generally introduces biases toward 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 ion map. It can be performed using sequential isolation windows (typically 10–50 Th) [53, 54] or with no isolation window [55–57] to generate the complex fragmentation spectra. However, the co-fragmentation of precursor ions leads to mixed product ions spectra, challenging data analysis and affecting selectivity [58, 59]. The elution and spectral matching against a reference spectral library are generally used to link the precursors to their product ions. Quantification is performed based on peak integration of the selected fragment ion chromatographic traces. The acquisition of a full product ion map of all present peptide ions allows reanalysis 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) [60] have become the gold standard for accurate quantitative analysis with greater sensitivity. More recently, parallel reaction monitoring (PRM) [61] was shown to dramatically improve the selectivity of measurements. Targeted proteomics consists in the monitoring of proteotypic peptide ions and a selected number of their product ions. It is generally used for precise and accurate protein quantification but has also been implemented for the measurement of a larger number of targeted peptides with less emphasis on the quantification accuracy. The latter has significantly increased the proteome coverage of targeted experiments [62]. Targeted protein quantification allows for higher sensitivity, wider dynamic range and greater reproducibility of measurements.

While shotgun approaches allow the greatest scale for analysis, their bias toward high abundance proteins could penalize the detection of splice variants that would generally be present in low abundance. DIA would circumvent this limitation as detection is not biased by abundance. However the spectral libraries needed for the unambiguous identification and quantification of these peptides are generally obtained using shotgun DDA data from data repositories thus having the same limitation. Targeted proteomics would allow to target the proteotypic peptides of isoforms. This renders a full quantitative approach using SRM or PRM possible [38]. Synthetic peptides can be used to optimize LC and MS parameters and as internal standards. These synthetic peptides also allow generating spectral libraries used for identification by spectral matching (for PRM and DIA).

3 Protein digestion and proteases in proteomics

Proteolytic cleavage after basic amino-acid residues is generally preferred to generate peptides in a bottom-up proteomics strategy. Peptides with a basic residue at their C-termini typically show an increase in the ionization efficiency during proton adduct electrospray ionization and yield fairly predictable fragmentation patterns. Such peptides would therefore include at a minimum two basic moieties, the N-terminal α-amino group and the guanidinium group (Arg) or ε-amino group (Lys) at their C-termini, and therefore at least two protonation sites. The three basic amino-acid residues present in proteins are lysine, arginine and histidine. To date, no protease is known to cleave at histidine residues. When the histidine residue is preceded by a threonine or a serine, cleavage at the histidine residue can be performed with low specificity

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

using copper II. The cleavage at other histidine sites is 10–100-fold slower [63]. 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. However, when a proline occurs at the C-terminus of Lys or Arg, the bond is almost completely resistant to trypsin. Lys-N has been described more recently and cleaves at the N-terminus of lysine residues [64] while LysargiNase cleaves at the N-terminus of both Arg and Lys residues [65]. Peptides generated by Lys-N and LysargiNase have shown a propensity to produce more b type ions in CID and c ions in ETD [66] when compared to enzymes that cleave at the C-terminus of basic residues. The amino acid composition of proteins varies depending on their function and localization. For example, lysine occurrences range from 6 to 8% in extracellular, nuclear and cytoplasmic proteins to only 4.4% in membrane proteins. Arginine mean occurrences range from 4 to 5% in all protein classes, except for nuclear proteins where it reaches 8.7% [67].

In contrast to basic amino acid residue 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 has a cleavage preference at the N-terminus of Asp with occasional cleavage at the N-terminus of Glu residues. Glu and Asp mean occurrences are around 5.5 and 5%, respectively [68]. Other enzymes cleave at aromatic or hydrophobic residues, which can be interesting for the analysis of membrane proteins. These enzymes include chymotrypsin and the less specific pepsin that cleave preferably at the C-terminus of Trp, Tyr, Phe and Leu. Recently, Meyer et al. described two proteases that cleave at aliphatic residues: wild-type α-lytic protease (WaLP) and its active site mutant (MaLP) [69]. Enzymes like SAP9 [70, 71] and OmpT [72] were introduced for extended bottom-up or middle-down proteomics as they generate larger peptides than the previously mentioned proteases. They cleave preferably at dibasic sites.

Some proteases are conformation specific and are used for particular digestions in native conditions such as papain [73] and IdeS [74,75] that cleave above and below the hinge region of immunoglobulin Gs (IgGs), respectively. Lys-C and pepsin are also known to cleave the hinge region of IgGs in nondenaturing buffer conditions. These proteases are frequently used for IgG and antibody drug conjugates characterization [76–78]. In general, the specificity and activity of proteases require particular buffer and pH conditions. In a recent review, Tsiatsiani and Heck discussed the different proteases employed in large scale proteomics experiments [42].

4 Complementarity and orthogonality

While trypsin allows the highest proteome coverage among all proteolytic enzymes, the trypsin non-accessible part of the proteome remains significant due to the uneven distribution of trypsin cleavage sites (lysine and arginine residues) across protein sequences resulting either in very short or very long peptides not suited for conventional LC-MS analyses. More specifically, peptides shorter than five to seven amino acids are mainly redundant and cannot be assigned to a unique protein sequence, while peptides larger than 5 kDa tend to have an adverse behavior in classical LC-MS settings and require high-resolution MS measurement for the correct assignment of precursor and product ions charge states. Furthermore, most commonly used database search engines are not efficient for the analysis of these large peptides that are generally highly charged (4+ and higher). For targeted protein quantification, protein-specific peptides in the 8–25 amino-acid residue range are generally selected [79]. However, the trypsin inaccessible sequences can be of considerable interest, especially in cases where particular isoforms or PTMs are of biological significance. In these scenarios, enzymes alternative to trypsin may be complementary to trypsin in accessing 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 commonly employed enzymes (Lys-C, Lys-N, Asp-N, Arg-C and Glu-C) was performed. Glu-C has a particular behavior as its specificity depends on digestion conditions. As Glu-C cleaves at a slower rate after Asp and Asp-N cleaves at a slower rate before Glu [80], partial cleavage is more often observed with these two enzymes. Figure 1 and Supporting Information Table SI-1 compare 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 Supporting Information Table SI2 and Supporting Information Fig. SI-1.

Peptides shorter than eight residues are mostly redundant regardless of the enzyme used. Roughly, only 20% (most of them including five to seven residues) are unique. However, trypsin and Asp-N produce twice the amount of these short uninformative peptides than enzymes that cleave only at one residue (Lys-C, Arg-C).

The primary interest of proteases alternative to trypsin is the accessibility to otherwise unreachable sequences of the proteome. Lys-C/N and Arg-C 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 acid residues. To estimate the capacity of these two categories of enzymes to access distinct areas of the proteome, the sequence coverage of the human proteome using peptides in the 8–25 residues range and three different enzymes was modeled (Fig. 2A).

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),

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

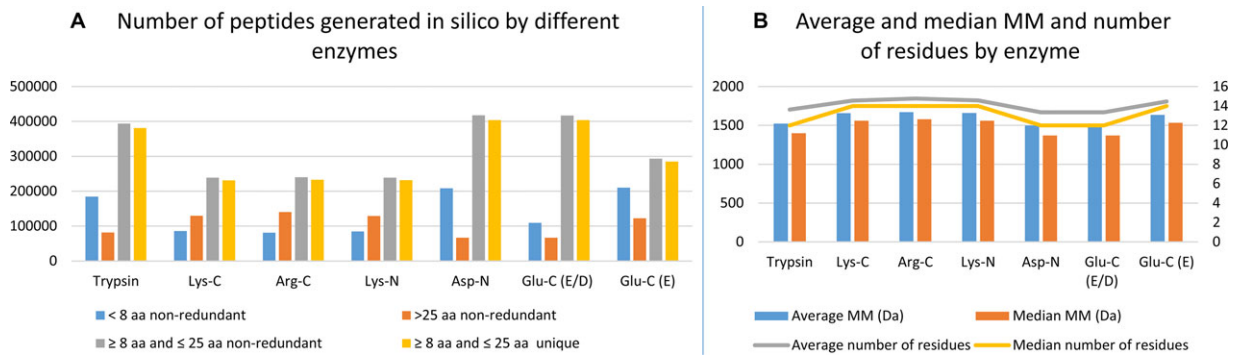


Figure 1. (A) Distribution and number of peptides generated by in silico digestion of the human proteome using seven different enzymes in the 8–25 amino acid (aa) residues range. (B) Average and median molecular mass (MM) (left Y axis) and average and median number of amino acid residues (right Y axis) in the 8–25 amino acid residues range.

whereas Lys-C can potentially add 7.5% of sequence coverage. This observation indicates a higher degree of orthogonality between Glu-C and trypsin than between 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 (see Materials and methods in Supporting Information). Searches were performed against a restricted database containing solely the UPS1 proteins using the search engine MASCOT. All 48 UPS1 proteins were identified for all evaluated digestion proteases. Only peptides without any missed-cleavage were used to determine protein coverages. The pie chart (Fig. 2B) 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 (62.5%) was obtained with the tryptic digestion while 57.9 and 41.5% of sequence coverage were achieved with Lys-C and Glu-C, respectively. The lower proteome coverage obtained with Lys-C and Glu-C is consistent with our simulation and other studies [43]. Moreover, the search engines for peptide identification are generally optimized for tryptic peptides and non-tryptic peptides tend to have lower scores [42]. By merging all the identification information obtained for the three enzymes, the global protein coverage of the UPS1 mixture rose to 81.1%. In this case, the use of alternative enzymes enabled to recover 49.5% of the sequence parts that were not covered by trypsin.

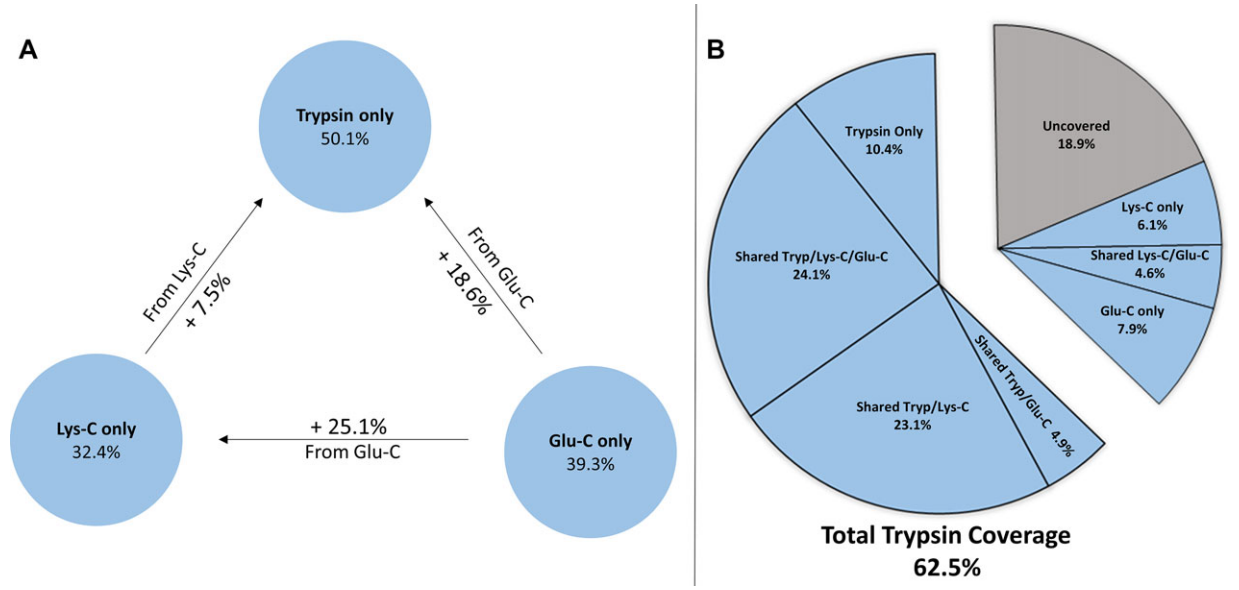


Figure 2. (A) Amino acid coverage of the human proteome with peptides in the 8–25 residues range using multiple in silico digestions. (B) Experimental amino acid coverage of a standard protein mixture (universal protein standard (UPS1)), using trypsin, Lys-C and Glu-C (E).

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

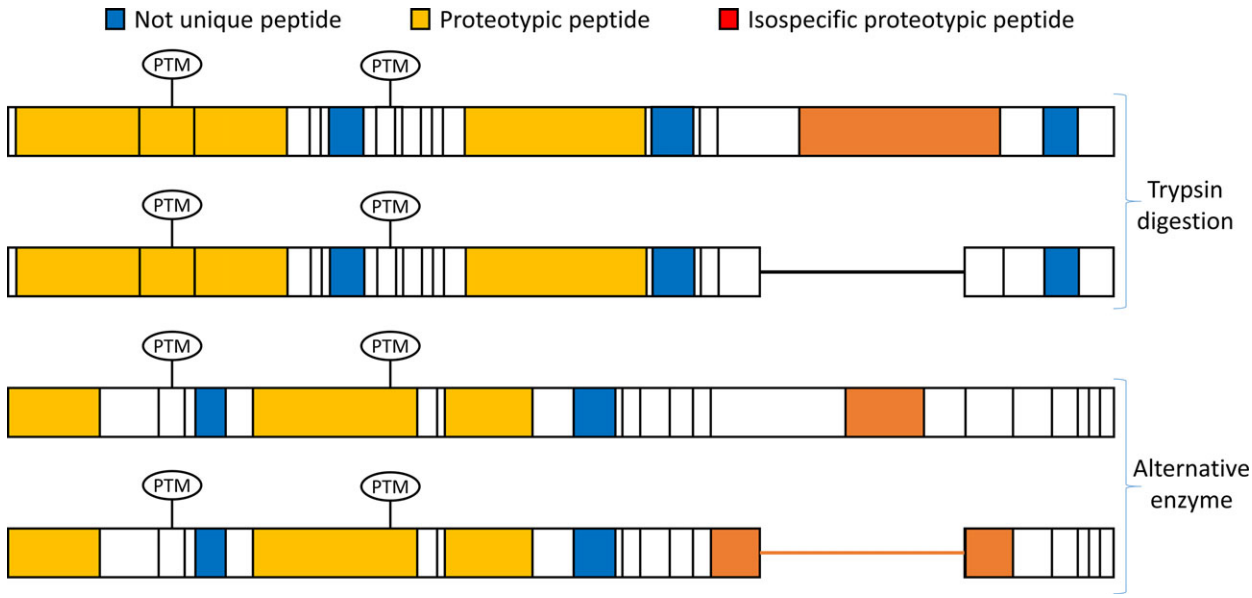


Figure 3. An illustration of complementarity provided by alternative digestion. Using trypsin alone, the second PTM and the splice variant would be missed. The alternative enzyme allows the identification of the splice variant junction and the second PTM but misses the first PTM and other parts of the sequence. Hence, using both enzymes gives access to complementary sequence information.

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 or to produce more proteotypic peptides to target more proteins. In a landmark study, Swaney et al. identified three times more unique peptides in yeast using multiple proteases compared to trypsin alone. This allowed the identification of 595 additional proteins as compared to just a trypsin digestion (3313 proteins) and the sequence coverage was also increased by nearly threefold [44]. Lesur et al. employed a Glu-C digestion to characterize at the peptide level the somatic EGF receptor's 746–750 deletion mutation [81]. In this particular case, tryptic proteolysis did not generate acceptable signature peptides for the unambiguous characterization of the mutation. The larger peptides produced by Lys-C and Arg-C have a better chance of spanning an intron as reported by Blakeley et al. (Arg-C and Lys-C can provide 27.2 and 25.5% of intron spanning peptides, respectively while trypsin and Glu-C only provide 20.4 and 17.9%) [38]. A 72% increase in the number of detected phosphopeptides was enabled by the parallel use of Lys-N and trypsin compared to trypsin alone [46]. For the characterization of protein N-termini, the presence of an arginine or a lysine near the N-terminal residue would prevent its identification. Alternative enzymes like Asp-N and chymotrypsin provided well-suited peptides for the identification of N-terminal heterogeneities in therapeutic monoclonal antibodies resulting from misprocessed signal peptide cleavage sites [82]. Asp-N in conjugation with trypsin allowed for the correction of erroneously predicted transit peptide cleavage sites of mitochondrial proteins [83]. In hydrogen/deuterium exchange approaches, trypsin is unsuited as a digestion enzyme since digestion has to be performed at low pH and temperature to inhibit the back exchange of deuterium to hydrogen; therefore pepsin is generally used [84]. Combining pepsin with other proteases active at acidic pH such as protease XIII from *Aspergillus saitoi* and protease XVIII from *Rhizopus* was suggested to increase sequence coverage [85, 86]. Nepenthesin from monkey cups (*Nepenthes*) has also been described as an alternative to pepsin in hydrogen/deuterium exchange experiments [87]. For deamidation assessment, Sap9 was found to be more reliable than trypsin as it cleaves faster at a lower pH preventing artifactual deamidations [71]. Figure 3 illustrates how an alternative enzyme can provide access to the sequences left behind by trypsin providing complementary proteotypic peptides to identify and quantify PTMs and sequence variants.

5 Technical considerations

5.1 Sample preparation

Trypsin digestion protocols have been optimized to ensure digestion efficiency and specificity. In general, the proteomics grade commercial trypsin is a modified version of the enzyme with limited autolytic activity that has also been treated to eliminate residual contaminating chymotrypsin activity. This modified trypsin has optimal activity in the pH range 7.8–8.7 and is resistant to up to 1 M urea. Several digestion buffers have been used, most often ammonium bicarbonate

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

or Tris-HCl. Prior to digestion, proteins are generally denatured in high concentrations of urea (up to 8 M), then disulfide bonds are reduced using dithiothreitol (DTT) or tris(2-carboxyethyl)phosphine (TCEP). The free thiol groups of cysteines are then alkylated using iodoacetamide. The urea concentration is subsequently reduced before adding the enzyme. Digestion is generally performed at 37°C. This protocol is, with some considerations, suitable for a number of alternative enzymes such as Lys-C, Lys-N, LysargiNase, Arg-C, Asp-N, Glu-C and chymotrypsin. The most significant particularities and considerations to be aware of are discussed in the following paragraph.

As stated earlier, trypsin cleavage is blocked at Arg or Lys residues located at the N-terminus of a proline. It is also slowed at multiple adjacent cleavage sites. Lys-C and Lys-N are more resistant to chemical denaturation and are active in urea concentrations exceeding 4 M and highly basic environments (up to pH 9.5). Lys-N is also tolerant to temperatures up to 70°C. Furthermore, digestion with Lys-C (and Arg-C) is not affected by the presence of proline. This has been exploited to enhance the generation of tryptic peptides: Lys-C digestion is often performed at high urea concentration (4–6 M), then urea is diluted to 0.8 M followed by trypsin digestion. This allows reducing the occurrence of miscleavages and increasing the digestion efficiency [88].

Arg-C is a cysteine protease active in reducing environments (presence of DTT, cysteine and calcium chloride). Oxidative agents and heavy metals inhibit its activity, hence EDTA is generally added to the digestion buffer. LysargiNase is most active at pH 7.5 and tolerant to temperatures up to 55°C. However, it is less tolerant to chaotropic agents and starts to lose activity at 0.3 M urea. TCEP reduction is preferred to DTT as the latter affects the efficiency of LysargiNase [65].

The activity and specificity of Glu-C are dependent on the pH and buffer used. This protease preferentially cleaves glutamyl bonds in ammonium acetate pH 4.0 or ammonium bicarbonate pH 7.8 whereas in phosphate buffers (pH 7.8) it cleaves both glutamyl and aspartyl bonds. AspN requires small quantities of zinc (0.5 mM zinc acetate) to enhance its activity. The α-lytic proteases WaLP and MaLP are less tolerant to urea than trypsin; therefore, sodium deoxycholate at 0.1% is preferred as chaotropic agent [69]. All enzymes do present non-specific and partial cleavages. Even with trypsin, which is reputed for offering the best cleavage specificity and completeness, miscleavages and non-specific cleavages are not uncommon [88–90]. This can be a source of bias in quantification experiments. Protein digestion is a major source of variability in peptide abundances. However, digestion performed in controlled conditions is reproducible and hence relative quantification remains achievable. Nevertheless, digestion specificity and completeness need to be controlled for absolute quantification. Isotope-labeled peptide concatemers and protein standards are to date the main tools that allow assessing and accounting for biases introduced by digestion [91–94].

5.2 Complexity reduction and impact on LC-MS ion density

As shown in Tables 1 and Supporting Information SI-1, trypsin generates nearly twice the amount of peptides compared to Lys-C, Lys-N or Arg-C and a significant number of those are shorter than five amino acid residues. Lys-C/N and Arg-C peptides are on average 1.45 and 1.51 times larger than tryptic peptides, respectively. In addition to increasing the chance of identifying combinations of PTMs, generating a smaller number of larger peptides can have beneficial effects during LC-MS analysis. The lower sample complexity is expected to translate into 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 interferences in targeted analysis, hence an increase in quantification accuracy. The hydrophobicity factors (HF) of the three different proteome digests were calculated and a bar chart of the number of peptides observed in different hydrophobicity factor bins was produced (Fig. 4). Trypsin, which generates the largest number of peptides, exhibits in the medium hydrophobicity range [12-38 HF factor] more than 1.6 times more peptides than Arg-C or Lys-C. Interestingly, the Lys-C and Arg-C digests are, in contrast to a general belief, only marginally enriched with hydrophobic peptides compared to trypsin digests, and 18.4% of Lys-C peptides and 17.6% of Arg-C peptides are highly hydrophobic compared to 14.8% of tryptic peptides. However, the overall number of hydrophobic peptides is higher for trypsin (93 482) when compared to Lys-C (69 225) and Arg-C (73 881) peptides.

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 (see Materials and methods in Supporting Information). The left panels of Fig. 5 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 most densely the space, especially in the 10–50 min range. Similarly to the results obtained by hydrophobicity factor calculations, the Lys-C digest, which contains larger peptides on average compared to trypsin, does not present a denser area at the end of the gradient which suggests that Lys-C does not significantly produce more hydrophobic peptides than trypsin. The panels on the right side of Fig. 5 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 the quadrupoles when

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

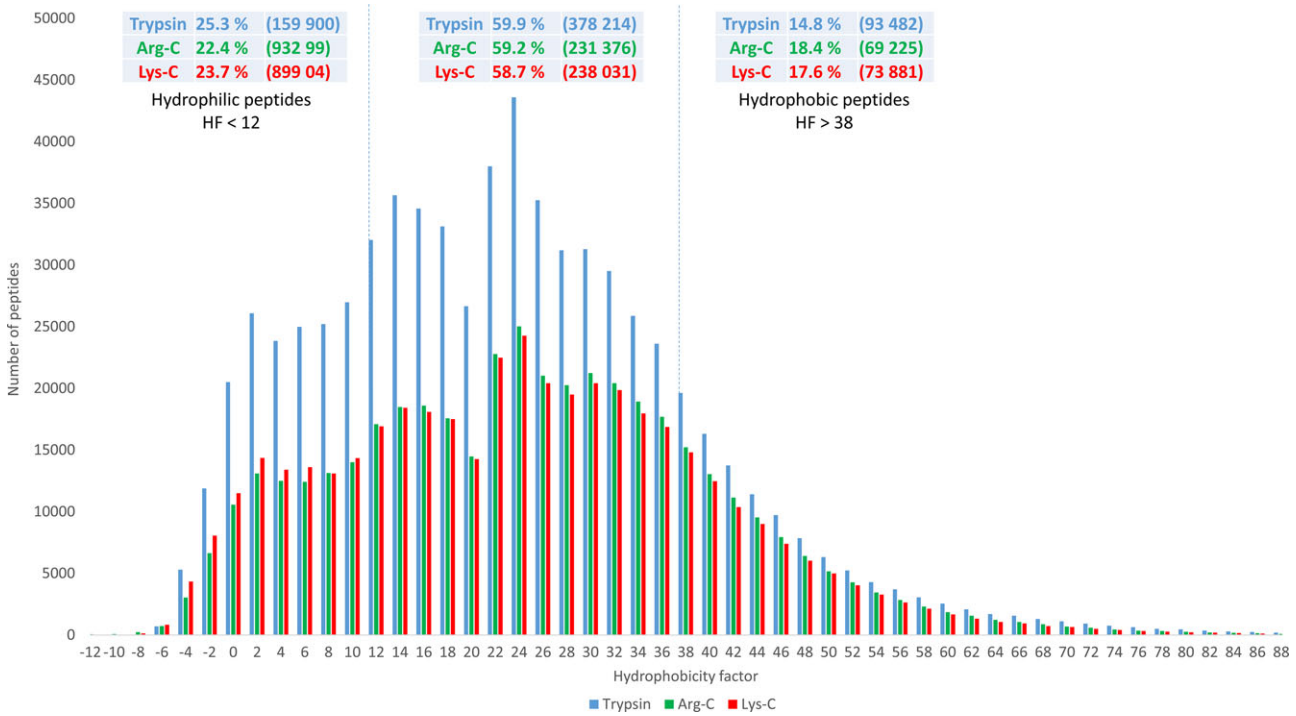


Figure 4. Hydrophobicity factors of all peptides in the 5 residues- 5 kDa range generated by trypsin, Arg-C and Lys-C in silico digestion of the human proteome and calculated using the SSRcalc algorithm [95, 96]. The inset tables show the percentage and number of peptides for each enzyme in three ranges: very hydrophilic (HF < 12), very hydrophobic (HF > 38) and the range in between.

targeting a specific peptide which may result in decreased signal interferences due to co-isolation.

5.3 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 [60] aiming at the evaluation of the influence of the parameters affecting the CID fragmentation of peptides in the collision cell (in triple quadrupole mass spectrometers or in the HCD cell of Orbitrap instruments) showed that the 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 at 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 [97], were generated by PRM analysis (see Materials and methods in Supporting Information). For "typical" tryptic peptides, i.e. doubly charged peptides comprising 10–16 amino acids, two main scenarios can generally be distinguished. In the first one, illustrated in Supporting Information Fig. SI–2A 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 Supporting Information Fig. SI-2B 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 (20.85 ev)) 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

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

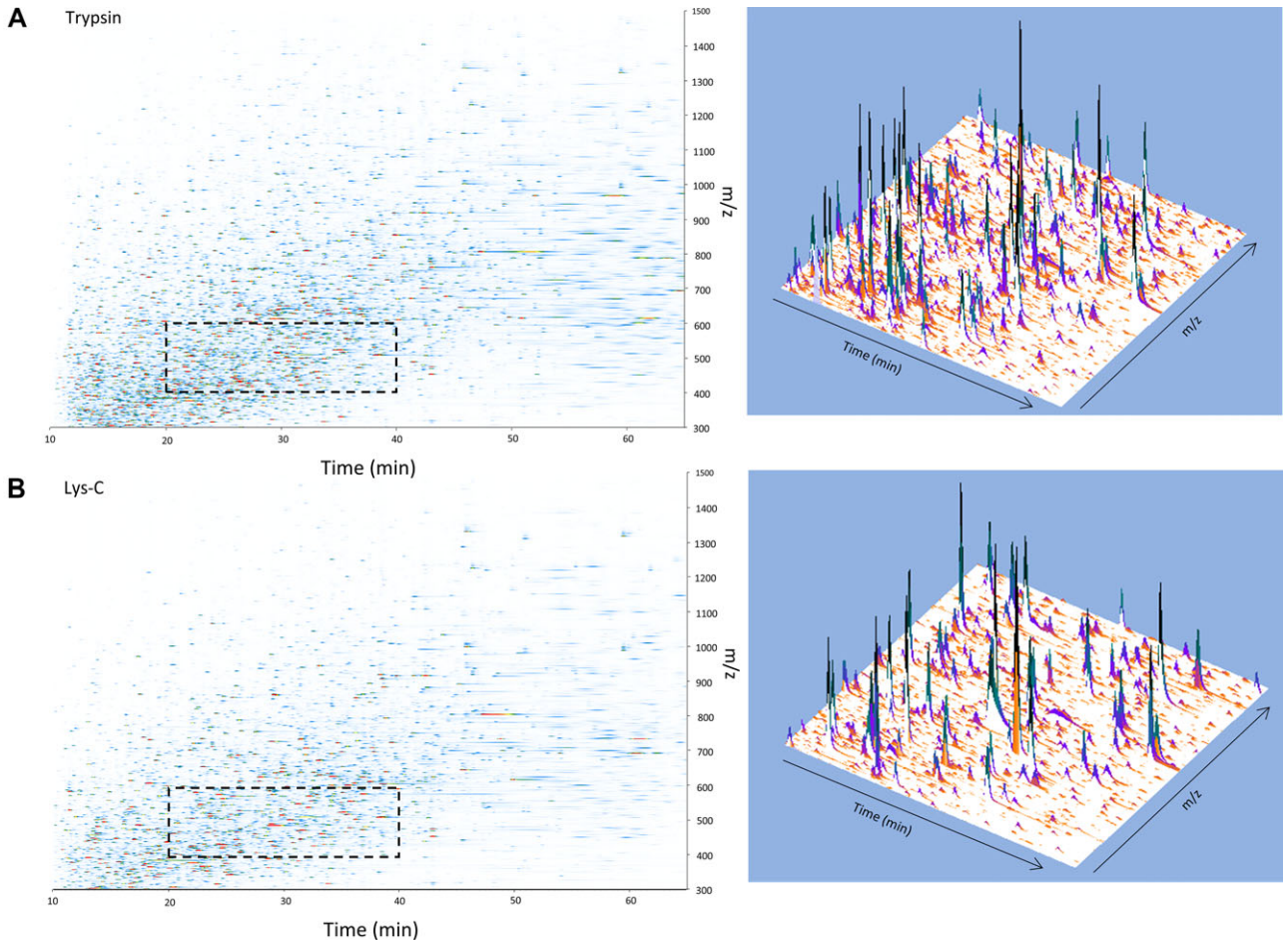


Figure 5. LC-MS heat map of depleted human plasma digested with trypsin (A) and Lys-C (B). The right panels are 3D representations of the peak density in the rectangles delimited by the dashed lines in the left panels.

proteins was performed and resulted in a similar observation. This is illustrated in Supporting Information Fig. SI-2C 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 fully 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 to provide the highest number of peptide identifications by conventional database searching algorithms [98–100]. Although it represents a simple approximation, and these values were primarily used for identification, i.e. generation of a wide fragmentation pattern, quantitative assays would benefit from fewer but 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 Fig. 6. Figure 6A represents the pseudo-breakdown curves of the peptide AIPVAQDLNAPSDWDSRGK (*m/z* 683.348, z = 3+) together with the MASCOT ion score as a 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+}) (Fig. 6B). It was 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

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

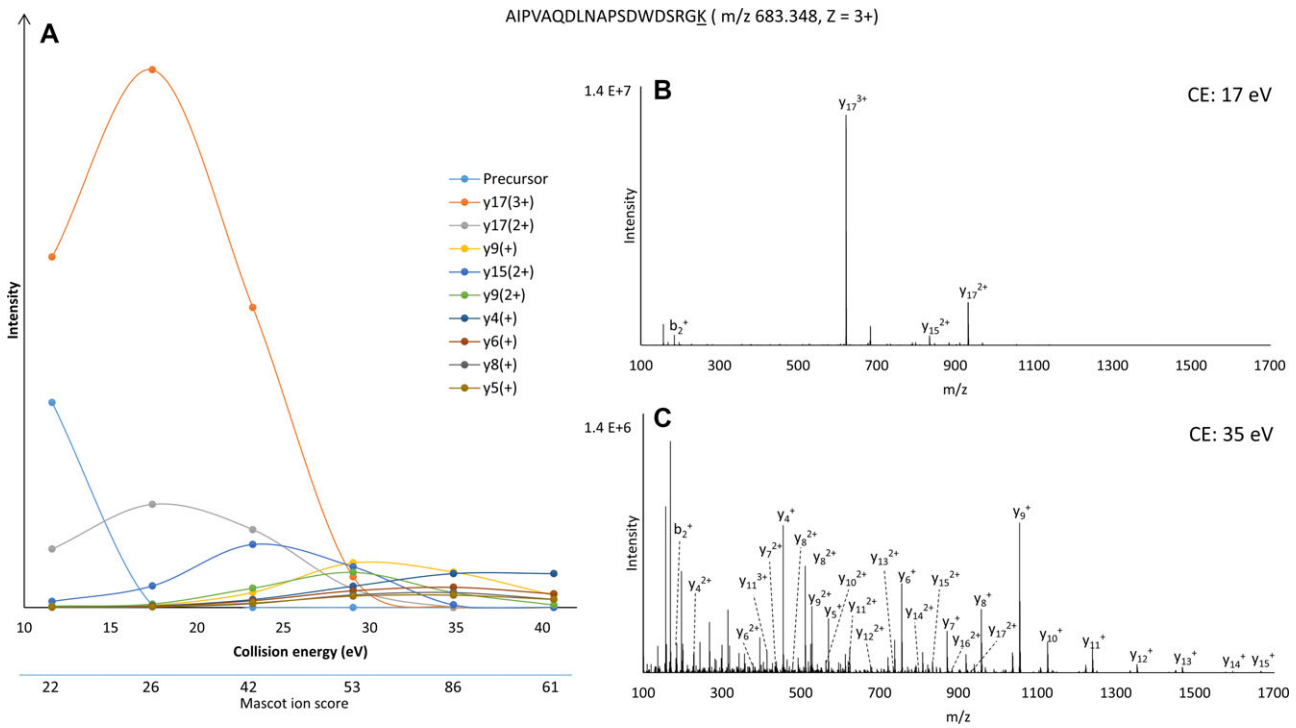


Figure 6. Pseudo breakdown curves of the AIPVAQDLNAPSDWDSRGK peptide. Fragment ion intensity in function of the collision energy applied and MASCOT ion score for each collision energy (A). MS2 spectrum at CE 17 ev (nCE 10) (B) and at CE 35 ev (nCE 30) (C).

increased number of fragments assigned, in spite of lower overall intensity, until a maximum of 86 is obtained for a CE of 35 ev (Fig. 6C).

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 above was based on their pseudo-breakdown curves. For each peptide, the intensity of the most intense fragment (base peak) ion across the six evaluated nCEs 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 Fig. 7A. Figure 7B shows that for the majority of these 159 peptides, the normalized collision energy generating the most intense MS/MS base peak is lower than the generally used nCE 25.

Such a peptide-specific optimization of collision energy results in a clear benefit for the sensitivity of measurement in PRM quantification experiments. 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 twofold was observed.

6 Conclusion

Proteomics approaches based on multiple protease digestions provide a more complete view of the proteome. The increased sequence coverage resulting from complementary digestions is in demand, as it allows accessing information-rich sequences lost by trypsin digestion. This is crucial when targeting specific proteoforms in which PTMs, single amino-acid substitutions or alternative splicing are not accessible using trypsin for identification as well as for quantification. Furthermore, enzymes cleaving at a single site (such as Lys-C or Arg-C) have a significant impact on ion density in LC-MS as illustrated with the digest of human plasma samples. Moreover, Lys-C and Arg-C provide more intron spanning unique peptides that help identifying splice variants. The almost exclusive use of trypsin in shotgun LC-MS approaches has resulted in a certain undercoverage of isoforms. Also, databases, search engines and other data analysis tools are almost exclusively trypsin-centric and tend to underperform with non-tryptic peptides. Moreover, only limited effort has been devoted to date toward building proteoform centric repositories, which would then allow systematic identification of isoforms. Targeted proteomics approaches are well suited to identify and quantify sequence variants and PTMs, by

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

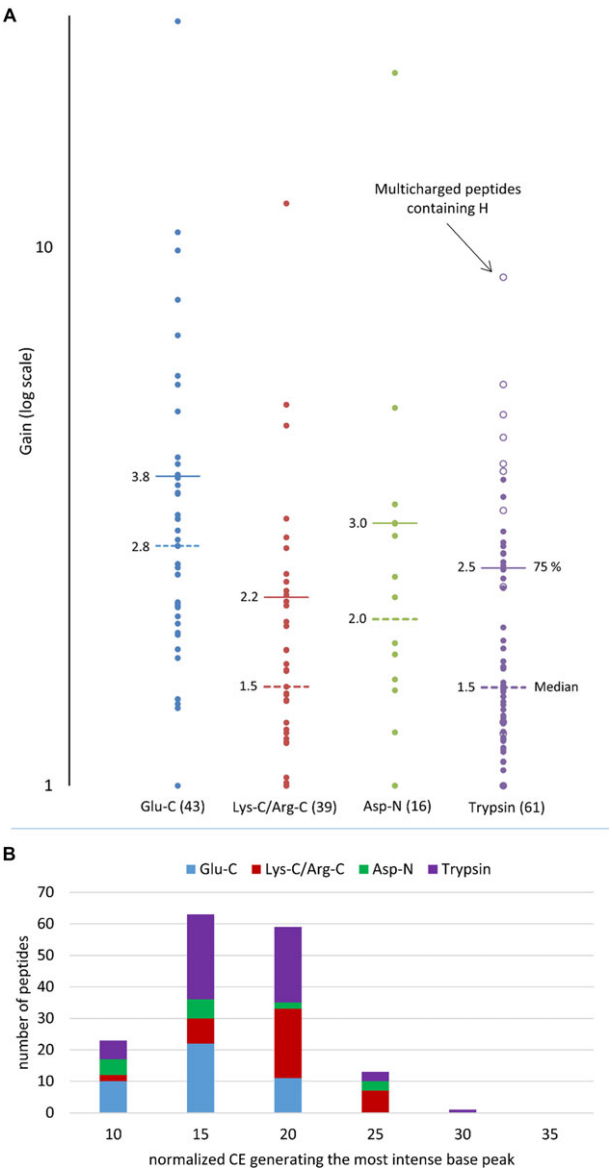


Figure 7. (A) Gain in sensitivity (log scale) of 159 peptides categorized by peptide structure. Based on the breakdown curves of each peptide, the gain was defined as the ratio of the intensity of the most intense fragment ion across the six evaluated nCEs compared to that measured at the "regular" normalized collision energy of 25 nCE. Dashed lines represent the median value and solid lines represent the upper quartile value. (B) Distribution of the optimum nCE for the 159 peptides.

targeting proteotypic peptides of the proteoforms of interest. The recent development of high-resolution MS platforms to perform quantitative studies allow analyzing isoforms and modifications with a higher degree of selectivity and sensitivity. The use of a default collision energy, even though providing higher identification scores during database searches, was shown to have limitations in product ion based quantification. Contrary to DIA, PRM allows using the optimal collision energy for each precursor. The sensitivity of targeted quantification assays can be improved by optimizing the collision energy to produce a few intense product ions rather than a high number of fragments as it is done for identification.

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) and Dr. Guy Berchem (CHL) for providing the plasma samples, Sang-Yoon Kim for technical assistance and Dr. Jan van Oostrum for helpful discussions.

The authors have declared no conflict of interest.

7 References

[1] Lander, E. S., The new genomics: global views of biology. *Science* 1996, *274*, 536–539.

[2] Pennisi, E., ENCODE project writes eulogy for junk DNA. *Science* 2012, *337*, 1159–1161.

[3] Pan, Q., Shai, O., Lee, L. J., Frey, B. J., Blencowe, B. J., Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 2008, *40*, 1413–1415.

[4] Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I. et al., Alternative isoform regulation in human tissue transcriptomes. *Nature* 2008, *456*, 470–476.

[5] Ahlf, D. R., Compton, P. D., Tran, J. C., Early, B. P. et al., Evaluation of the compact high-field orbitrap for top-down proteomics of human cells. *J. Proteome Res.* 2012, *11*, 4308–4314.

[6] Sheynkman, G. M., Shortreed, M. R., Frey, B. L., Smith, L. M., Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. *Mol. Cell. Proteomics* 2013, *12*, 2341–2353.

[7] Smith, L. M., Kelleher, N. L., Proteoform: a single term describing protein complexity. *Nat. Meth.* 2013, *10*, 186–187.

[8] Nilsen, T. W., Graveley, B. R., Expansion of the eukaryotic proteome by alternative splicing. *Nature* 2010, *463*, 457–463.

[9] Nern, A., Nguyen, L. V., Herman, T., Prakash, S. et al., An isoform-specific allele of Drosophila N-cadherin disrupts a late step of R7 targeting. *Proc. Natl. Acad. Sci. USA* 2005, *102*, 12944–12949.

[10] Brinkman, B. M. N., Splice variants as cancer biomarkers. *Clin. Biochem.* 2004, *37*, 584–594.

[11] Yamazaki, T., Wälchli, S., Fujita, T., Ryser, S. et al., Splice variants of Enigma homolog, differentially expressed during heart development, promote or prevent hypertrophy. *Cardiovasc. Res.* 2010, *86*, 374–382.

[12] Hansson, O., Zhou, Y., Renström, E., Osmark, P., Molecular function of TCF7L2: consequences of TCF7L2 splicing for molecular function and risk for type 2 diabetes. *Curr. Diab. Rep.* 2010, *10*, 444–451.

[13] Gabut, M., Samavarchi-Tehrani, P., Wang, X., Slobodeniuc, V. et al., An alternative splicing switch regulates embryonic

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

stem cell pluripotency and reprogramming. *Cell*, *147*, 132–146.

[14] Velloso, C. P., Harridge, S. D. R., Insulin-like growth factor-I E peptides: implications for ageing skeletal muscle. *Scand. J. Med. Sci Sports* 2010, *20*, 20–27.

[15] Wei, J., Zaika, E., Zaika, A., p53 family: role of protein isoforms in human cancer. *J. Nucleic Acids* 2012, *2012*, 687359.

[16] Webb, K. E., Martin, J. F., Hamsten, A., Eriksson, P. et al., Polymorphisms in the thrombopoietin gene are associated with risk of myocardial infarction at a young age. *Atherosclerosis*, *154*, 703–711.

[17] Bates, D. O., Harper, S. J., Therapeutic potential of inhibitory VEGF splice variants. *Future Oncol.* 2005, *1*, 467–473.

[18] Kelleher, N. L., Thomas, P. M., Ntai, I., Compton, P. D., LeDuc, R. D., Deep and quantitative top-down proteomics in clinical and translational research. *Expert Rev. Proteomics* 2014, *11*, 649–651.

[19] Domon, B., Aebersold, R., Options and considerations when selecting a quantitative proteomics strategy. *Nat. Biotechnol.* 2010, *28*, 710–721.

[20] Stastna, M., Van Eyk, J. E., Analysis of protein isoforms: can we do it better? *Proteomics* 2012, *12*, 2937–2948.

[21] Zhang, Z., Wu, S., Stenoien, D. L., Pasa-Tolic, L., High-throughput proteomics. *Annu. Rev. Anal. Chem. (Palo Alto Calif.)* 2014, *7*, 427–454.

[22] Cannon, J. R., Cammarata, M. B., Robotham, S. A., Cotham, V. C. et al., Ultraviolet photodissociation for characterization of whole proteins on a chromatographic time scale. *Analyt. Chem.* 2014, *86*, 2185–2192.

[23] Michalski, A., Damoc, E., Lange, O., Denisov, E. et al., Ultra high resolution linear ion trap orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Mol. Cell. Proteomics* 2012, *11*, O111.013698.

[24] Cammarata, M. B., Thyer, R., Rosenberg, J., Ellington, A., Brodbelt, J. S., Structural characterization of dihydrofolate reductase complexes by top-down ultraviolet photodissociation mass spectrometry. *J. Am. Chem. Soc.* 2015, *137*, 9128–9135.

[25] Mao, Y., Valeja, S. G., Rouse, J. C., Hendrickson, C. L., Marshall, A. G., Top-down structural analysis of an intact monoclonal antibody by electron capture dissociation-Fourier transform ion cyclotron resonance-mass spectrometry. *Anal. Chem.* 2013, *85*, 4239–4246.

[26] Riley, N. M., Westphall, M. S., Coon, J. J., Activated ion electron transfer dissociation for improved fragmentation of intact proteins. *Anal. Chem.* 2015, *87*, 7109–7116.

[27] Sarbu, M., Ghiulai, R., Zamfir, A., Recent developments and applications of electron transfer dissociation mass spectrometry in proteomics. *Amino Acids* 2014, *46*, 1625–1634.

[28] Brunner, A. M., Lössl, P., Liu, F., Huguet, R. et al., Benchmarking multiple fragmentation methods on an Orbitrap fusion for top-down phospho-proteoform characterization. *Anal. Chem.* 2015, *87*, 4152–4158.

[29] Fornelli, L., Damoc, E., Thomas, P. M., Kelleher, N. L. et al., Analysis of intact monoclonal antibody IgG1 by electron transfer dissociation Orbitrap FTMS. *Mol. Cell. Proteomics* 2012, *11*, 1758–1767.

[30] Tsybin, Y. O., Fornelli, L., Stoermer, C., Luebeck, M. et al., Structural analysis of intact monoclonal antibodies by electron transfer dissociation mass spectrometry. *Anal. Chem.* 2011, *83*, 8919–8927.

[31] Catherman, A. D., Durbin, K. R., Ahlf, D. R., Early, B. P. et al., Large-scale top-down proteomics of the human proteome: membrane proteins, mitochondria, and senescence. *Mol. Cell. Proteomics* 2013, *12*, 3465–3473.

[32] Dang, X., Scotcher, J., Wu, S., Chu, R. K. et al., The first pilot project of the consortium for top-down proteomics: A status report. *Proteomics* 2014, *14*, 1130–1140.

[33] Kellie, J. F., Tran, J. C., Lee, J. E., Ahlf, D. R. et al., The emerging process of Top Down mass spectrometry for protein analysis: biomarkers, protein-therapeutics, and achieving high throughput. *Mol. Biosyst.* 2010, *6*, 1532–1539.

[34] Tran, J. C., Zamdborg, L., Ahlf, D. R., Lee, J. E. et al., Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* 2011, *480*, 254–258.

[35] Richards, A. L., Hebert, A. S., Ulbrich, A., Bailey, D. J. et al., One-hour proteome analysis in yeast. *Nat. Protocols* 2015, *10*, 701–714.

[36] Beck, S., Michalski, A., Raether, O., Lubeck, M. et al., The impact II, a very high-resolution quadrupole time-of-flight instrument (QTOF) for deep shotgun proteomics. *Mol. Cell. Proteomics* 2015, *14*, 2014–2029.

[37] Khatun, J., Yu, Y., Wrobel, J. A., Risk, B. A. et al., Whole human genome proteogenomic mapping for ENCODE cell line data: identifying protein-coding regions. *BMC Genom.* 2013, *14*, 141.

[38] Blakeley, P., Siepen, J. A., Lawless, C., Hubbard, S. J., Investigating protein isoforms via proteomics: a feasibility study. *Proteomics* 2010, *10*, 1127–1140.

[39] Nesvizhskii, A. I., Proteogenomics: concepts, applications and computational strategies. *Nat. Meth.* 2014, *11*, 1114–1125.

[40] Wang, X., Liu, Q., Zhang, B., Leveraging the complementary nature of RNA-Seq and shotgun proteomics data. *Proteomics* 2014, *14*, 2676–2687.

[41] Hartmann, E. M., Armengaud, J., N-terminomics and proteogenomics, getting off to a good start. *Proteomics* 2014, *14*, 2637–2646.

[42] Tsiatsiani, L., Heck, A. J., Proteomics beyond trypsin. *FEBS J.* 2015, *282*, 2612–2626.

[43] 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*, 1573–1584.

[44] 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*, 1323–1329.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

[45] Giannone, R. J., Wurch, L. L., Podar, M., Hettich, R. L., Rescuing those left behind: recovering and characterizing underdigested membrane and hydrophobic proteins to enhance proteome measurement depth. *Anal. Chem.* 2015, *87*, 7720–7728.

[46] Gauci, S., Helbig, A. O., Slijper, M., Krijgsveld, J. et al., Lys-N and trypsin cover complementary parts of the phosphoproteome in a refined SCX-based approach. *Anal. Chem.* 2009, *81*, 4493–4501.

[47] Wiśniewski, J. R., Mann, M., Consecutive proteolytic digestion in an enzyme reactor increases depth of proteomic and phosphoproteomic analysis. *Anal. Chem.* 2012, *84*, 2631–2637.

[48] Domon, B., Aebersold, R., Mass spectrometry and protein analysis. *Science* 2006, *312*, 212–217.

[49] Aebersold, R., Mann, M., Mass spectrometry-based proteomics. *Nature* 2003, *422*, 198–207.

[50] Tabb, D. L., Vega-Montoto, L., Rudnick, P. A., Variyath, A. M. et al., Repeatability and reproducibility in proteomic identifications by liquid chromatography–tandem mass spectrometry. *J. Proteome Res.* 2010, *9*, 761–776.

[51] Paulovich, A. G., Billheimer, D., Ham, A.-J. L., Vega-Montoto, L. et al., Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS Platform Performance. *Mol. Cell. Proteomics* 2010, *9*, 242–254.

[52] Bell, A. W., Deutsch, E. W., Au, C. E., Kearney, R. E. et al., A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat. Meth.* 2009, *6*, 423–430.

[53] Gillet, L. C., Navarro, P., Tate, S., Röst, H. et al., Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics* 2012, *11*, O111.016717.

[54] Lesur, A., Domon, B., Advances in high-resolution accurate mass spectrometry application to targeted proteomics. *Proteomics* 2015, *15*, 880–890.

[55] 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*, 144–156.

[56] Geromanos, S. J., Vissers, J. P., Silva, J. C., Dorschel, C. A. et al., 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*, 1683–1695.

[57] Li, G. Z., Vissers, J. P., Silva, J. C., Golick, D. et al., 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*, 1696–1719.

[58] Gallien, S., Duriez, E., Demeure, K., Domon, B., Selectivity of LC-MS/MS analysis: implication for proteomics experiments. *J. Proteomics* 2013, *81*, 148–158.

[59] 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. *Mol. Cell. Proteomics* 2015, *14*, 1411–1418.

[60] Gallien, S., Duriez, E., Domon, B., Selected reaction monitoring applied to proteomics. *J. Mass Spectrom.* 2011, *46*, 298–312.

[61] Gallien, S., Domon, B., Quantitative proteomics using the high resolution accurate mass capabilities of the quadrupole-orbitrap mass spectrometer. *Bioanalysis* 2014, *6*, 2159–2170.

[62] Gallien, S., Kim, S. Y., Domon, B., Large-scale targeted proteomics using internal standard triggered-parallel reaction monitoring (IS-PRM). *Mol. Cell. Proteomics* 2015, *14*, 1630–1644.

[63] Allen, G., Campbell, R. O., Specific cleavage of histidine-containing peptides by copper (II). *Int. J. Pept. Protein Res.* 1996, *48*, 265–273.

[64] Boersema, P. J., Taouatas, N., Altelaar, A. F. M., Gouw, J. W. et al., Straightforward and de novo peptide sequencing by MALDI-MS/MS using a Lys-N metalloendopeptidase. *Mol. Cell. Proteomics* 2009, *8*, 650–660.

[65] Huesgen, P. F., Lange, P. F., Rogers, L. D., Solis, N. et al., LysargiNase mirrors trypsin for protein C-terminal and methylation-site identification. *Nat. Methods* 2015, *12*, 55–58.

[66] Taouatas, N., Drugan, M. M., Heck, A. J. R., Mohammed, S., Straightforward ladder sequencing of peptides using a Lys-N metalloendopeptidase. *Nat. Meth.* 2008, *5*, 405–407.

[67] 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*, 594–600.

[68] 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*, 5558–5569.

[69] Meyer, J. G., Kim, S., Maltby, D. A., Ghassemian, M. et al., Expanding proteome coverage with orthogonal-specificity alpha-lytic proteases. *Mol. Cell. Proteomics* 2014, *13*, 823–835.

[70] Laskay, U. A., Srzentic, K., Monod, M., Tsybin, Y. O., Extended bottom-up proteomics with secreted aspartic protease Sap9. *J. Proteomics* 2014, *110*, 20–31.

[71] Srzentic, K., Fornelli, L., Laskay, U. A., Monod, M. et al., Advantages of extended bottom-up proteomics using Sap9 for analysis of monoclonal antibodies. *Anal. Chem.* 2014, *86*, 9945–9953.

[72] Wu, C., Tran, J. C., Zamdborg, L., Durbin, K. R. et al., A protease for 'middle-down' proteomics. *Nat. Meth.* 2012, *9*, 822–824.

[73] Yan, B., Valliere-Douglass, J., Brady, L., Steen, S. et al., Analysis of post-translational modifications in recombinant monoclonal antibody IgG1 by reversed-phase liquid chromatography/mass spectrometry. *J. Chromatogr. A* 2007, *1164*, 153–161.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

[74] Fornelli, L., Ayoub, D., Aizikov, K., Beck, A., Tsybin, Y. O., Middle-down analysis of monoclonal antibodies with electron transfer dissociation Orbitrap Fourier Transform mass spectrometry. *Anal. Chem.* 2014, *86*, 3005–3012.

[75] Ayoub, D., Jabs, W., Resemann, A., Evers, W. et al., Correct primary structure assessment and extensive glyco-profiling of cetuximab by a combination of intact, middle-up, middle-down and bottom-up ESI and MALDI mass spectrometry techniques. *mAbs* 2013, *5*, 699–710.

[76] Beck, A., Wagner-Rousset, E., Ayoub, D., Van Dorsselaer, A., Sanglier-Cianferani, S., Characterization of therapeutic antibodies and related products. *Anal. Chem.* 2013, *85*, 715–736.

[77] Beck, A., Diemer, H., Ayoub, D., Debaene, F. et al., Analytical characterization of biosimilar antibodies and Fc-fusion proteins. *TrAC Trends Anal. Chem.* 2013, *48*, 81–95.

[78] Wagner-Rousset, E., Janin-Bussat, M. C., Colas, O., Excoffier, M. et al., Antibody-drug conjugate model fast characterization by LC-MS following IdeS proteolytic digestion. *mAbs* 2014, *6*, 173–184.

[79] 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. *Mol. Cell. Proteomics* 2015, *14*, 2331–2340.

[80] 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*, 2833–2844.

[81] Lesur, A., Ancheva, L., Kim, Y. J., Berchem, G. et al., Screening protein isoforms predictive for cancer using immunoaffinity capture and fast LC-MS in PRM mode. *PROTEOMICS – Clin. Appl.* 2015, *9*, 695–705.

[82] Ayoub, D., Bertaccini, D., Diemer, H., Wagner-Rousset, E. et al., Characterization of the N-terminal heterogeneities of monoclonal antibodies using in-gel charge derivatization of alpha-amines and LC-MS/MS. *Anal.I Chem.* 2015, *87*, 3784–3790.

[83] Vaca Jacome, A. S., Rabilloud, T., Schaeffer-Reiss, C., Rompais, M. et al., N-terminome analysis of the human mitochondrial proteome. *Proteomics* 2015, *15*, 2519–2524.

[84] Zhang, X., Less is more: membrane protein digestion beyond urea-trypsin solution for next-level proteomics. *Mol. Cell. Proteomics* 2015, *14*, 2441–2453.

[85] Cravello, L., Lascoux, D., Forest, E., Use of different proteases working in acidic conditions to improve sequence coverage and resolution in hydrogen/deuterium exchange of large proteins. *Rapid Comm. Mass Spectrom.* 2003, *17*, 2387–2393.

[86] Zhang, H. M., Kazazic, S., Schaub, T. M., Tipton, J. D. et al., Enhanced digestion efficiency, peptide ionization efficiency, and sequence resolution for protein hydrogen/deuterium exchange monitored by Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* 2008, *80*, 9034–9041.

[87] Rey, M., Yang, M., Burns, K. M., Yu, Y. et al., Nepenthesin from monkey cups for hydrogen/deuterium exchange mass spectrometry. *Mol. Cell. Proteomics* 2013, *12*, 464–472.

[88] Glatter, T., Ludwig, C., Ahrné, E., Aebersold, R. et al., 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*, 5145–5156.

[89] Fang, P., Liu, M., Xue, Y., Yao, J. et al., Controlling nonspecific trypsin cleavages in LC-MS/MS-based shotgun proteomics using optimized experimental conditions. *Analyst* 2015, *140*, 7613–7621.

[90] Picotti, P., Aebersold, R., Domon, B., The implications of proteolytic background for shotgun proteomics. *Mol. Cell. Proteomics* 2007, *6*, 1589–1598.

[91] Brun, V., Dupuis, A., Adrait, A., Marcellin, M. et al., Isotope-labeled protein standards: toward absolute quantitative proteomics. *Mol. Cell. Proteomics* 2007, *6*, 2139–2149.

[92] Brun, V., Masselon, C., Garin, J., Dupuis, A., Isotope dilution strategies for absolute quantitative proteomics. *J. Proteomics* 2009, *72*, 740–749.

[93] Scott, K. B., Turko, I. V., Phinney, K. W., Quantitative performance of internal standard platforms for absolute protein quantification using multiple reaction monitoring-mass spectrometry. *Anal. Chem.* 2015, *87*, 4429–4435.

[94] Pratt, J. M., Simpson, D. M., Doherty, M. K., Rivers, J. et al., Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes. *Nat. Protocols* 2006, *1*, 1029–1043.

[95] Krokhin, O. V., Spicer, V., Peptide retention standards and hydrophobicity indexes in reversed-phase high-performance liquid chromatography of peptides. *Anal. Chem.* 2009, *81*, 9522–9530.

[96] Krokhin, O. V., Spicer, V., Predicting peptide retention times for proteomics. *Curr. Protoc. Bioinformatics* 2010, *Chapter 13*, Unit 13.14.

[97] Kim, Y. J., Sertamo, K., Pierrard, M. A., Mesmin, C. et al., Verification of the biomarker candidates for non-small-cell lung cancer using a targeted proteomics approach. *J. Proteome Res.* 2015, *14*, 1412–1419.

[98] Kelstrup, C. D., Jersie-Christensen, R. R., Batth, T. S., Arrey, T. N. et al., Rapid and deep proteomes by faster sequencing on a benchtop quadrupole ultra-high-field Orbitrap Mass spectrometer. *J. Proteome Res.* 2014, *13*, 6187–6195.

[99] Scheltema, R. A., Hauschild, J.-P., Lange, O., Hornburg, D. et al., The Q Exactive HF, a benchtop mass spectrometer with a pre-filter, high performance quadrupole and an ultra-high field Orbitrap analyzer. *Mol. Cell. Proteomics* 2014, *13*, 3618–3708.

[100] Michalski, A., Damoc, E., Hauschild, J.-P., Lange, O. et al., Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell. Proteomics* 2011, *10*, M111.011015.

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim